Evolving Affective Abstract Art through Measures Learned from a Corpus of Human-Made Art

Jules Verdijk
Graduation Thesis
Media Technology MSc program, Leiden University
August 2015
Thesis advisors: *Peter van der Putten* and *Eelco den Heijer*

Abstract. In this paper a new approach in evolutionary art is taken by evolving affective abstract art based on measures learned from a corpus of human-made art. Through a web survey 200 artworks were rated on the affect dimensions pleasure, arousal and dominance, and 69 image features were extracted from each artwork. Regression models were learned based on these features and the rated affect and tested on their accuracy in predicting affect, resulting in an affect measure for each dimension. Due to low consistency in the affect ratings and restrictions in the used image features, the measures had low accuracy in predicting arousal and dominance. Thus, the study focussed on using the measure for pleasure as the fitness function in a novel evolutionary system that generates art within a restricted style, including arousal and dominance as secondary measures. Sets of artworks were evolved for specific goals of affect, of which 90 were selected for a web survey to validate the affect of the images produced by the system. Results of this survey showed that the system is capable of generating art on a wide range of affect, but is only effective in generating art for affect goals on the extremities of the pleasure, arousal and dominance dimensions.

1. Introduction

The main goal of the field of evolutionary art is to produce aesthetically pleasing and novel artwork. A major problem in reaching that goal is how to measure aesthetics and novelty and how to use those measurements as a way to evolve images, music or other forms of art. In this paper a different approach is taken to evolve art. Instead of generating images with high aesthetic value, the focus will be on evolving abstract art that induces certain emotions, or more precise, evolving affective abstract images. Using affect as the goal when evolving abstract art is a novel approach and this study researches whether affect is an interesting and feasible addition to, or alternative for, aesthetic goals. The primary goal of this research was to evolve affective art based on a corpus of existing art. This was done through the development of a full end-to-end system. First a set of existing human-made abstract art was rated on three affect dimensions. The resulting database of images with affect ratings were then analysed on their features, and through machine learning measures for affect were extracted. These measures were then used as a fitness function in a novel evolutionary system that generates abstract images based on a limited set of possible shapes. Finally the effectiveness of the whole system to evolve affective abstract images was tested in a small user-test. A secondary goal of this study was to explore this process and find the opportunities and limitations of directly applying learned measures as fitness functions to evolve art.

The remainder of this paper is structured around several experiments that were ran to gain insight on different aspects of the development of the evolutionary system. First a short background is provided on evolutionary art and affect in section 2. Section 3 describes how the affect measures were learned,

including a web survey in which existing abstract art was rated on affect, the extraction of image features from this art and experiments on machine-learning measures for the affect of abstract art. This resulted in a set of affect measures of which the use in a novel evolutionary art system is described in section 4. The effectiveness of this system to generate art within a broad range of affect is described, after which the process and result of evolving art for different affect goals is presented. This section on evolving art concludes with the results of a second web survey, used to validate whether the system was successful in generating affective art. The results and possibilities of using learned affect measures to evolve art are then discussed in section 5, followed by an overall conclusion in section 6.

2. Background

An important part in the development of a system to evolve affective art is how to measure the affect of art. This section first provides an overview of commonly used methods in determining measures when evolving art for aesthetic value. Then a short description is given of what affect is, how it can be represented and how it is currently used in generating art, concluding with a comparison of existing research on the relation between image features and the affect of (abstract) images.

2.1. Goals of evolutionary art

There have been many attempts to evolve art that has aesthetic value. In general three main approaches are used to evaluate the aesthetic fitness in these evolutionary setups. Commonly used is interactive evolutionary computation, in which (in)direct subjective human evaluation is used as the fitness function [9,32]. For example the work by Draves, that applies online user votes as a basis for a fitness function to evolve screensavers [5]. Another approach is using aesthetic measures that are based on art theory. For example the work by den Heijer, who uses existing theories on aesthetics, like fractal dimensions and image complexity, to construct fitness functions [14,15]. An alternative method is retrieving aesthetic measures from examples of art. This is the corpus-based approach, in which a dataset of (human-made) art is used as a basis to learn aesthetic measures that can be applied as fitness functions [16]. This method is for example used to generate music in the work by Phon-Amnuaisuk et al., in which features learned from existing examples of music are applied to rate the fitness of newly generated music [23]. So far this approach has not directly been used for the purpose of evolving abstract images.
Combinations of these approaches are also used, like the work by Baluja et al., in which fitness functions are automated based on the user preferences of an initial evaluation round [1]. But none of these approaches has of yet solved the problem of computational aesthetic evaluation, the problem of providing an evolutionary system with the ability to judge the aesthetics of its generated artefacts [10]. McCormack considers that this is due to the fact that human aesthetic evaluation is not yet fully understood and that aesthetic evaluation cannot always directly be translated to computable measures [19]. Other than aesthetic values, evolutionary art may also have the goal of evolving novel art [4,8].

2.2. Affect

Affect representations. Evolving art with the goal of inducing emotion is less studied. Emotion itself however, is well researched within psychology and is usually referred to as affect, the general feeling or experience of emotion. Affect can be seen as a more general abstraction of emotion, mood, attitude and feelings and is often described as a state of affect, which can be represented in different forms. Often used forms are the dimensional and categorical representations. The dimensional, or factor-based, representation defines affective states through a value on one or more dimensions. Like Russel's pleasure-displeasure and arousal model [29] or Mehrabian's pleasure, arousal, dominance

model [20], in which affective states are represented in three values ranging from -1 to 1. Categorical representations define affective states by assigning them to set categories, like Ekman's six basic emotions (happiness, sadness, anger, surprise, disgust and fear) or Plutchik's wheel of emotions [7,24]. The used representation depends on what is studied or measured, categorical emotions are often used to report on or describe the emotion that people feel, while the dimensional representation is used to define any sort of affect, being it emotion, mood or attitude towards a person or object [3].

Generating affective art. In the field of generative art, affect has mainly been as used to produce affective music. Like the work by Scirea et al. [30], in which affective music is generated on the pleasure and arousal dimensions based on knowledge of the relation between musical features and mood, or the work by Birchfield, that uses theory on emotion to evolve qualitative music [2]. However, as of now no studies have been conducted on using measures for affect learned from a corpus of art to evolve abstract images.

Table 1. Overview of studies on affective abstract image classification and the resulting best scoring feature types, including the findings of this paper.

Study	Affect representation	Dataset	Ratings	Feature types used	Best scoring feature types
Machajdik, Hanbury 2010	Categorical	228 abstract paintings*	14 per painting	Colour, texture, composition, content	Features based on art theory.
Zhang et al, 2011	exciting-boring and relaxing-irritating	100 abstract images	20 per image	Features from biological image classification	Colour, edges and texture
Yanulevskaya et al, 2012	Positive-negative	500 abstract paintings	20 per painting	LAB-colours and SIFT-descriptions downgraded to bag-of-word descriptions	LAB-colours.
Wang et al, 2013	Categorical	228 abstract paintings*	14 per painting	Figure-ground relationship, colour pattern, shape, composition	Not included
Zhao et al, 2014	Categorical	228 abstract paintings*	14 per painting	Six artistic principles: Balance, emphasis, harmony, variety, gradation, movement	Emphasis, variety and symmetry, harmony for positive emotions
This paper	Dimensional (PAD)	200 abstract images	4 to 18 per image	Colour, texture, edges, shapes, canvas	Colour, texture, shapes

* = similar dataset

Affective image classification. Several studies have researched feature-based inference of affect and emotion in abstract artworks, using different datasets, affect representations and classification methods. Because of these differences in methodical setup it is difficult to compare their results in accurately predicting the affect of images. Comparable however, are the image features used to predict and classify affect. In general these features can be divided into image features based on art theory and low-level image features, where the last group of features is mostly used in studies that also studied secondary sets of images, like natural images from the International Affective Picture System (IAPS).

Table 1 gives an overview of the best scoring features found in relevant studies. Overall the consensus is that high-level features based on artistic principles and art theory are more effective in classifying the categorical affect of images than features based on low-level features [18,37].

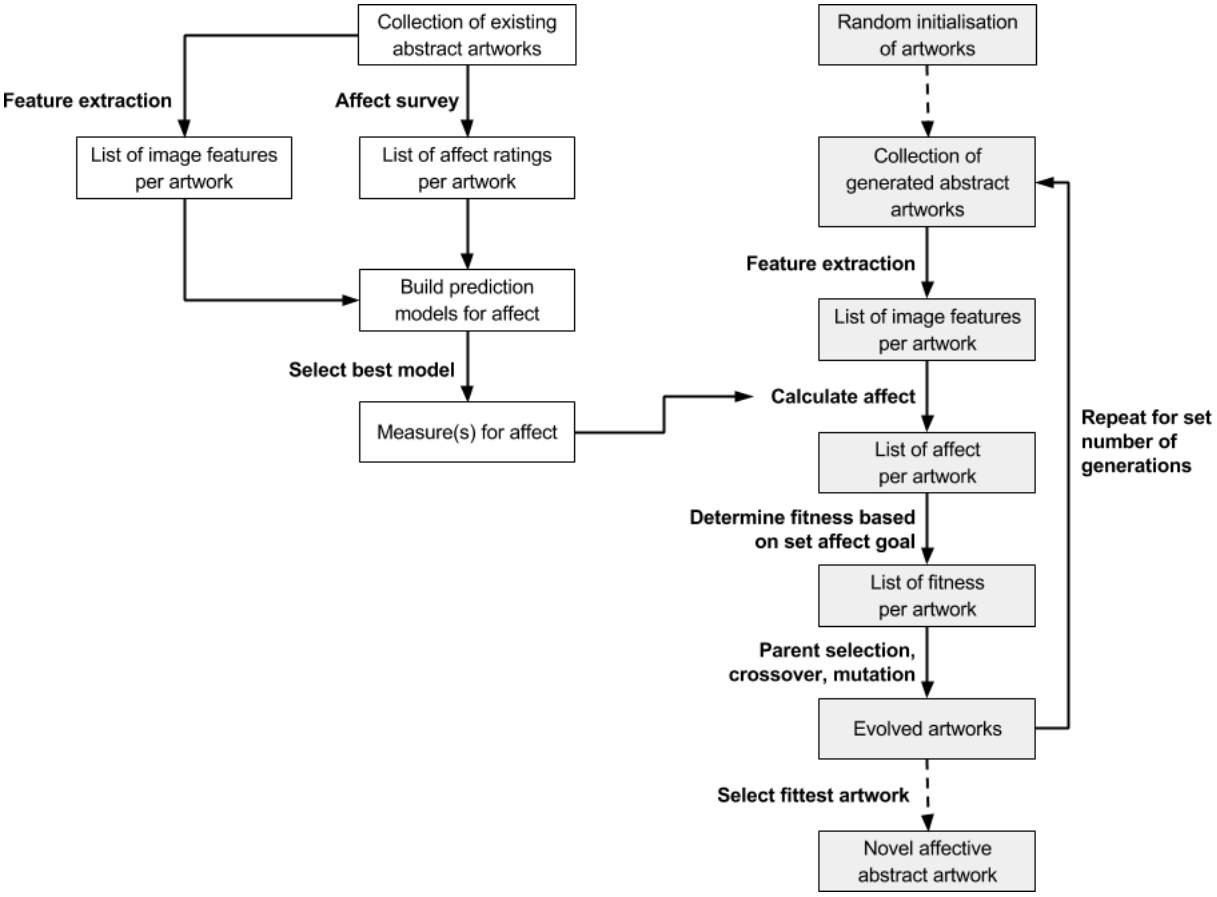

Figure 1. System flow for generating affective art based on existing abstract art.

3. Learning affect measures

To study the use of a corpus-based approach to evolve affective art a full end-to-end system was developed. This section presents the left part of the system as seen in figure 1, describing how affect measures were learned from a corpus of abstract art. First a representation of affect was chosen based on prior research. A set of abstract artworks was retrieved online and rated on three affect dimensions through a web survey, after which image features were computed from each artwork. Machine learning was then used to build prediction models based on the image features and the affect ratings. The resulting models, as well as models using a limited set of features, were tested on their prediction accuracy and analysed on their suitability as fitness functions. Finally a set of models was chosen to be used as affect measures.

3.1. Affect representation

As affect can be represented in different forms, the chosen form is essential for this research as it will influence the scoring of affect, the way models are learned and the way fitness is represented in the evolutionary system. For this study Mehrabian's pleasure, arousal and dominance (P, A, D) representation was chosen [20]. In this model, affect is represented through three continuous values ranging from -1 to 1, one value for pleasure, one for arousal and one for dominance. This

representation was chosen because prior studies on affective image classification used either a categorical representation [18,22] or the two-dimensional valence, arousal representation [17]. Using the additional dominance dimension may yield new insights on affective image classification. Also, affect is represented on the PAD-dimensions through three continuous values, making it directly suitable for computation and providing the option to evolve for very specific affect. So, instead of evolving "happy" art, an aim could be 0.6 pleasure, 0.1 arousal and -1.0 dominance, which offers a broader range of possibilities in exploring affect as a measure for evolutionary art. Another reason for using this representation was the availability of the AffectButton [3], an interactive self-report method for affect on the PAD-dimensions that is easily implementable in an online environment.

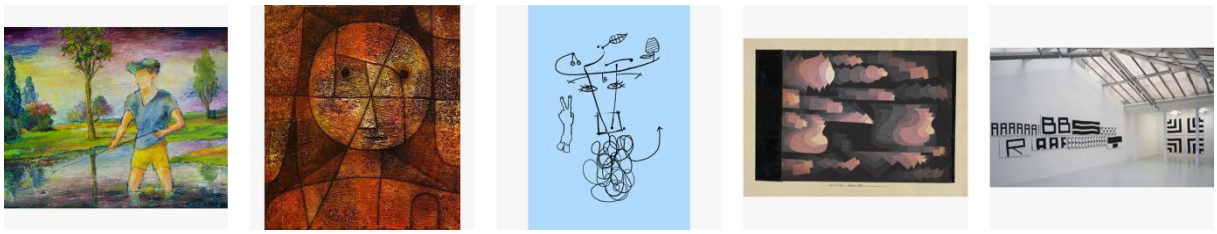

Figure 2. Examples of artworks that were rejected because they contain recognizable objects (landscape, person, face), image frames or walls.

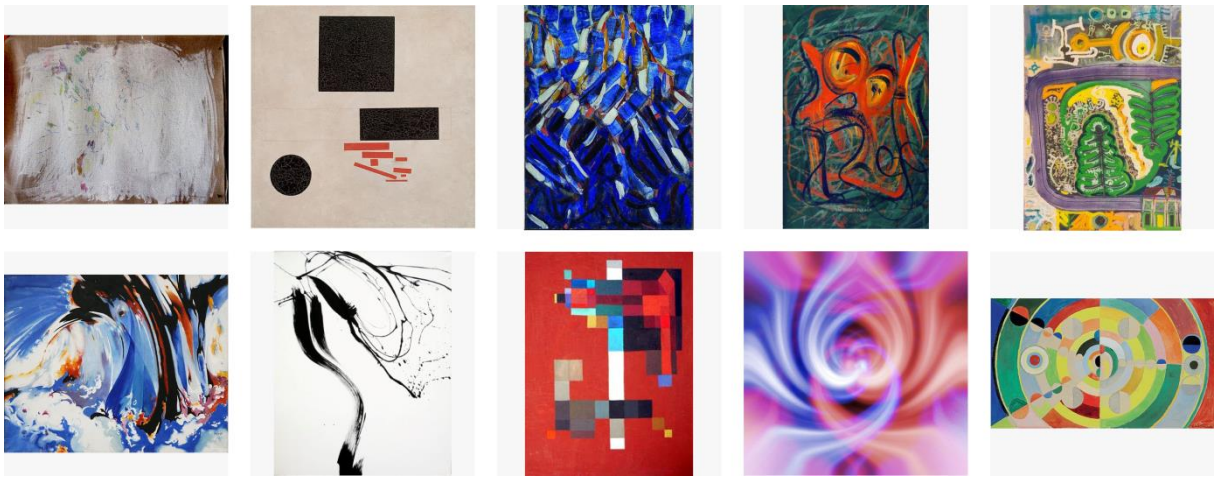

Figure 3. Examples of artworks selected for the corpus.

3.3. Corpus of affect rated abstract art

Original artworks. A total of 419 abstract artworks was retrieved from Wikimedia Commons in a .jpg format [1]. Because the evolutionary system is only capable of generating shapes and colours, the set of artworks was manually filtered to create a corpus of examples that is similar to the possible output of the system. Examples of rejected and selected artworks can be found in figure 2 and 3. Artworks were removed from the dataset based on the following aspects:

- Contained frames, walls or other signs of photography.
- Contained recognizable objects like faces, landscapes or text.
- Showed signs of compression or other distortion.
- Duplicates or remakes of other artworks in the dataset.
- Not a minimal width or height of 500 pixels.

[1] http://commons.wikimedia.org/wiki/Category:Abstract_paintings, retrieved at 10-6-2015

Of the remaining images, 200 were selected at random to serve as the corpus for learning affect measures. The images were resized to a maximum width or height of 500 pixels to be suitable to be displayed on a website, the resized images were considered to be the definite dataset and used in that form for feature extraction.

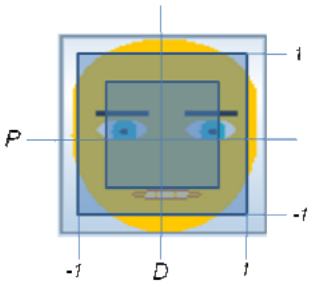

Figure 4. Mapping of the 2D-coordinates to the PAD dimensions, A is mapped to the distance to the center square. From Broekens, 2009 [3].

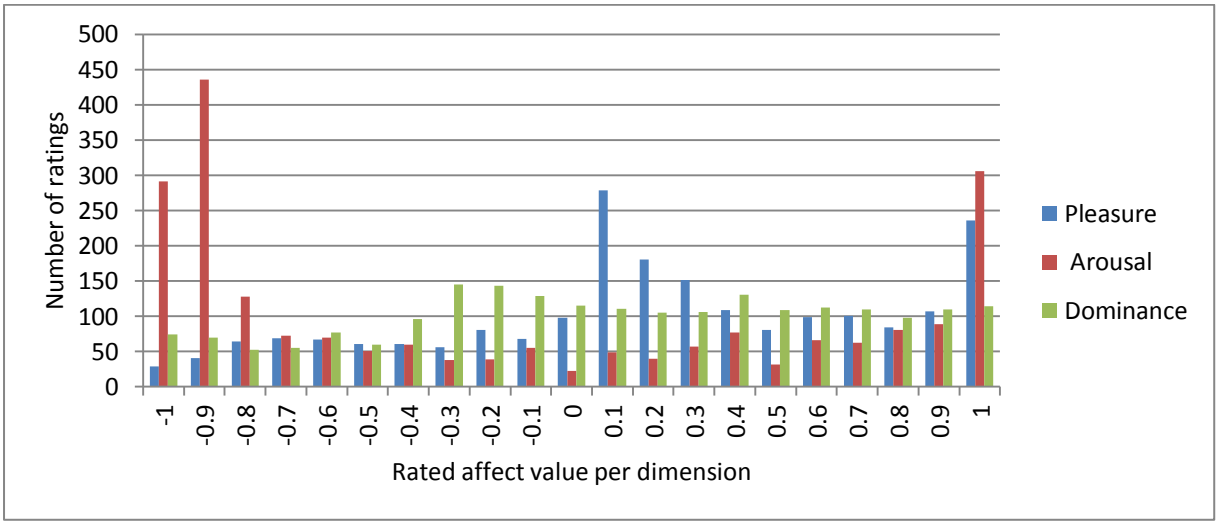

Figure 5. Histograms of all human affect ratings per dimension.

Online affective feedback. To gather data on the affective properties of the abstract artworks, a web survey was created where participants could rate the artworks on its affect with the AffectButton, which can be seen in figure 4 [3]. The AffectButton is an interactive self-report method for gathering affective feedback on the PAD-dimensions and is easily implementable in an online environment. There are however limitations to the AffectButton. The first is that participants need a short period of training to explore the full affect range of the button. A second limitation is that due to the two-dimensional interface, the rated arousal value is a derivative of the pleasure and dominance values. This has the consequence that the reliability of the ratings on the arousal dimension are relatively low, showing a bias towards the extremities -1 and 1. Although no study was done on using the AffectButton to rate the affect of art, the reliability of the P and D dimensions was determined to be good for rating the affect of a variety of samples in studies by Broekens [3].

Before the test, participants were given a short introduction regarding the subject of the research and an explanation of the AffectButton as provided by Broekens. They were asked to turn off any form of music or radio and given the instruction to "find the face that best matches the emotion that you think the artwork communicates". This instruction was chosen to have participants reflect more on the visual properties of the artworks, instead of on their internal emotion. Participants were given an unlimited amount of time to test the AffectButton without stimuli during the instructions. After stating that they were done exploring the AffectButton, 25 abstract artworks were sequentially shown,

selected at random from the total set of 200 images. This low number of ratings per participant was chosen to minimize human fatigue [32]. After giving feedback on an artwork, a delay of 2 seconds was implemented before showing the next artwork, during which the AffectButton was disabled and no image was shown. This delay was implemented to minimize the influence of the affect of an artwork on the rating of subsequent artwork. The affective feedback was stored in the form of three values (P,A,D) per rating, ranging from -1 to 1 with a decimal precision of 6.

Results. A total of 85 participants were recruited online ranging in age between 20 and 78 (M = 37.25, SD = 15.05; 51 men, 34 women). Because the goal of this study is creating affective abstract art in general, age and sex where considered to be irrelevant for the subsequent parts of this study. Each participant rated 25 abstract artworks, resulting in 2125 unique ratings. The rated artworks were randomly selected, so not all artworks received the same number of ratings (M=10.63, SD = 3.10), the most frequently rated artwork received 18 ratings, the least frequently rated received 4 ratings. As was to be expected, the PAD-dimensions were not rated independent of each other [3], a correlation between P and A ($r(2125) = 0.20$, $p < 0.001$) and P and D ($r(2125) = 0.17$, $p < 0.001$) was found. But no significant correlation was found between A and D, which is surprising because the A value directly depends on P and D (see figure 4). The mean ratings per dimension were 0.154 for P (SD = 0.55), -0.204 for A (SD = 0.78) and 0.041 for D (SD = 0.56). As can be seen in figure 5, the ratings for P and D were quite evenly spread from -1 to 1, with a peak around 1 and 0.1 for P and a slight preference for positive D values. The ratings for A however were clearly biased towards the extremities -1 and 1, which is possibly caused by the lower reliability of A due to the mentioned setup of the AffectButton. Other than ratings per dimension, it is also useful to look at ratings per artwork, for which the mean rating per artwork and its standard deviation were calculated. The mean values were used to build an interactive 3D-visualisation based on the ratings of the artworks on the three affect dimensions, a screenshot of this can be seen in Figure 6. This visualization, which clearly showed the correlation between P and A and between P and D, was used to get insight on the actual relation between the visual appearance of an artwork and its rating on the three dimensions. These insights helped in determining which features to extract (see section 3.4). As a measure for the consistency in the affect ratings, the standard deviation of the mean rating per dimension per artwork was used. The mean standard deviation per artwork for all dimensions was 0.598, P had the lowest mean standard deviation per artwork of 0.469, A the highest with a mean standard deviation of 0.717 and D had a mean standard deviation per artwork of 0.511. A correlation was found between the mean standard deviation per artwork for all dimensions and the number of ratings per artwork ($r(200) = 0.23$, $p < 0.001$).

3.4. Feature extraction

Based on the results of previous studies (table 1) it is to be expected that high-level image features based on art theory and artistic principles are more suitable to be used as measures for affect than low-level image features. As a secondary indication for possible image features that are related with affect, the visualisation as seen in figure 6 was used. The most clear differences in appearance are seen between artworks scoring high on all dimensions and artworks scoring low on all dimensions. The largest distinction is in the vividness of the colours and the amount of different colours used. Less distinctive, but still noticeable, are the differences in the entropy and the texture of the artworks. The results of previous studies and the insights from the visualisation were used to determine the set of image features to compute. However, due to practical limitations only a limited set of high-level features was extracted, complemented with a set of low-level features. In total 69 features, divided into 8 feature types, were computed for each image in the dataset of 200 abstract artworks. These features are summarized in table 2 and the process used to extract the features can be seen in figure 7.

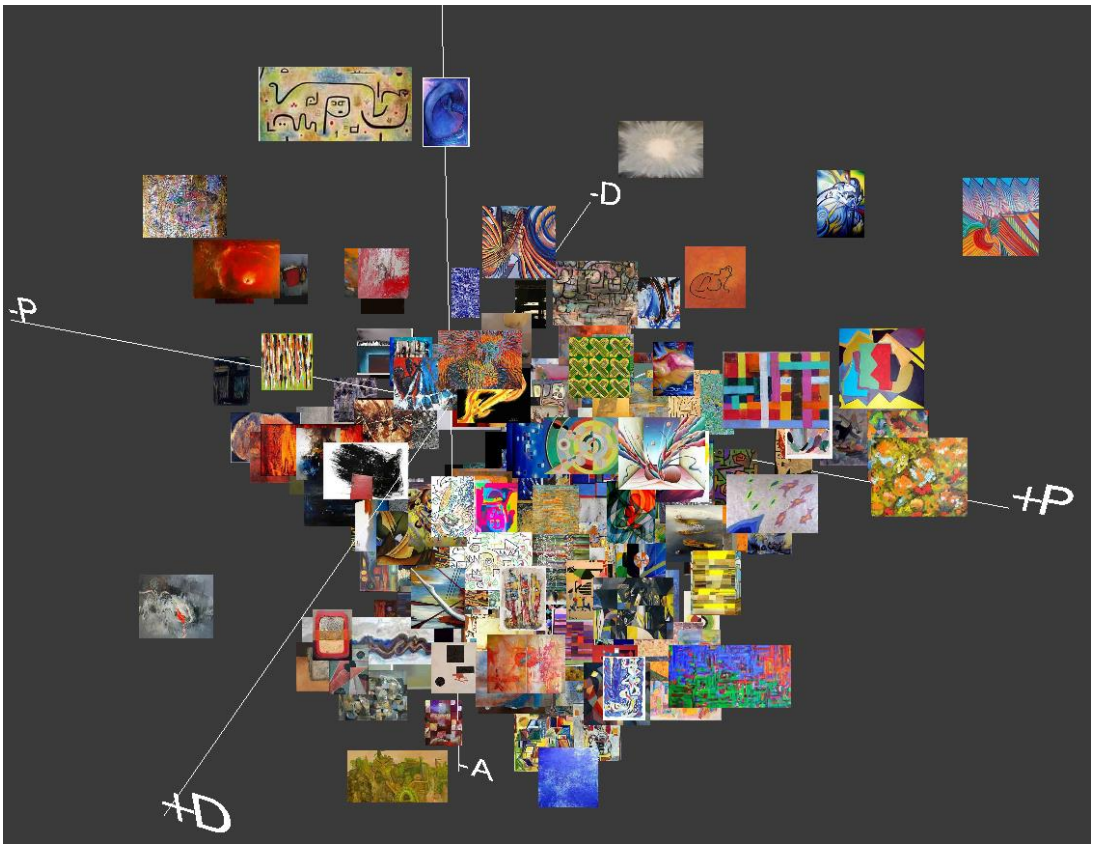

Figure 6. Example of the 3D-visualisation of the artworks placed in the PAD-dimension based on their mean affect rating.

Table 2. Summary of all extracted image features.

Feature type	#	Short Description
General properties	5	General properties of the image regarding size and aspect ratio.
Color	21	Mean values and standard deviations of the histograms of the Red, Green, Blue (RGB) color-spaces and the Hue, Saturation, Brightness (HSB) color-spaces and several custom nominal and numerical color descriptors.
Gray-scale measures	8	Statistical measures on the gray values of the 8-bit gray-scale conversion of an image.
Texture measures	5	Statistical measures on the Gray-Level Co-Occurrence Matrix (GLCM) [13,33].
Edges	2	Gray values of the image after running a Sobel edge detector [6].
Laplacian measures	5	Statistical measures on the gray values of the Laplacian transform of an image [21].
Binary image	2	Statistical measures on the gray values of a binary conversion of an image through an automatic threshold calculation based on the isoData algorithm [26].
Shape and region	25	Measures on the amount of shapes, properties of the largest shape (including size and position) [33], mean properties and their standard deviations for all found shapes [28] and the number of local maxima found.

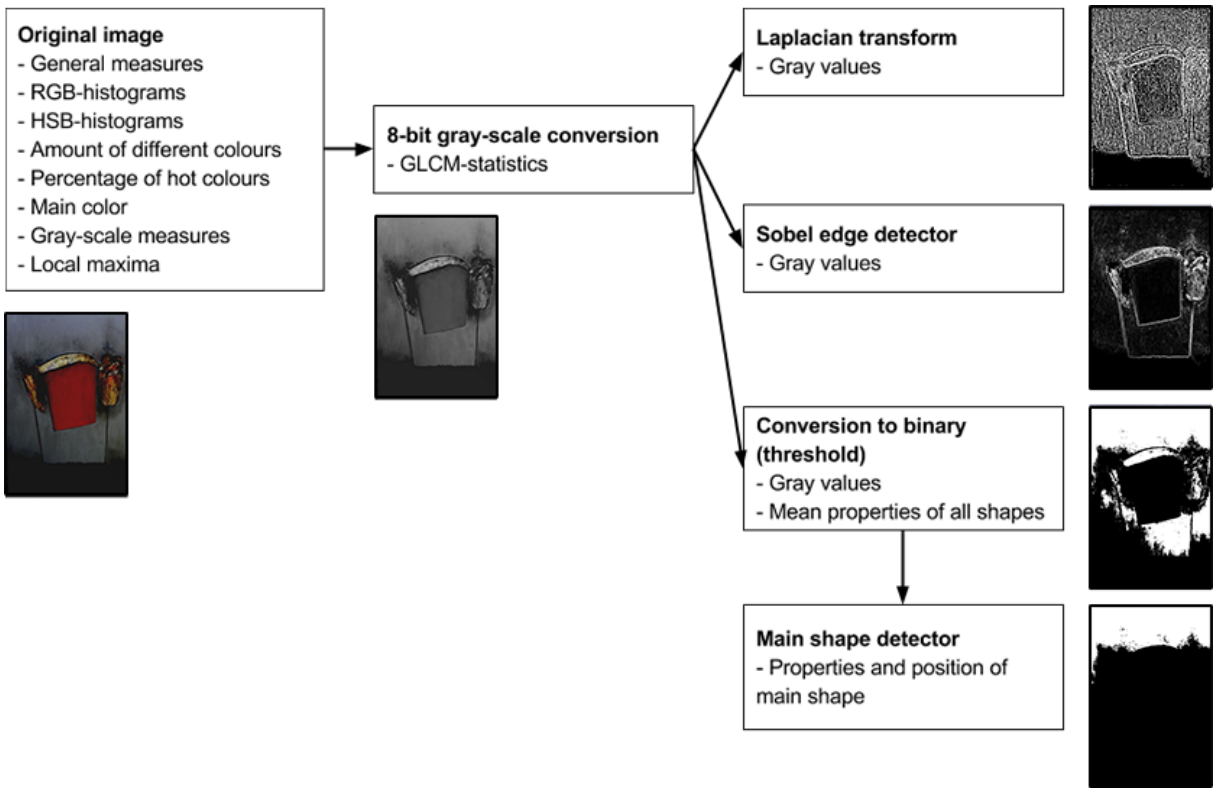

Figure 7. Processing and feature extraction of an image from the original dataset.

3.5. Prediction models for affect

The extracted features and the affective ratings were used to build prediction models for each of the PAD-dimensions separately. Preliminary experiments were done to determine whether pre-processing the dataset would be beneficial for the prediction accuracy. This included converting the ratings per dimensions to a binary problem, combining these binary ratings per dimension to eight total classes, only using the lowest (< -0.5) and highest (>0.5) rated images and using the mean affect ratings per artwork instead of all ratings. Preliminary results showed that this pre-processing would not be beneficial for the prediction accuracy. The result is that there are three separate regression problems, instead of one classification problem as seen in other studies. Because the models build for these problems should later serve as (combined) fitness functions in the evolutionary system, they were not only judged on their accuracy in predicting the affect values, but also on their suitability to serve as a fitness function. Requirements for this were that the models should be interpretable as well as use diverse feature-types to accommodate for a larger diversity in the optimization process of the evolutionary system. Diversity in used features was determined through the amount of feature-types and total features used.

Experimental setup. Two experiments were ran per affect dimension, using the algorithms and environment provided by the Weka experimenter [12]. The first experiment tested four regression algorithms on their accuracy and their suitability as fitness functions:

- ZeroR. Uses no features for classification, predicts the mean value. Used as a baseline comparison for the other classification algorithms.
- Linear Regression (LR) with M5 attribute selection method and elimination of collinear attributes.

- SMOreg with RegSMOImproved [31]. Implementation of the support vector machine for regression.
- M5Rules [25,34]. Builds a decision tree based on M5 and converts the best leafs into rules, where each rule contains a linear regression function.

The resulting models were tested on their prediction accuracy through ten repeated holdout runs with a randomized 66% training/test split and analysed on their suitability as fitness functions by looking at their interpretability and the number of used image features.
For the second experiment attribute selection algorithms were tested for the best scoring regression model. This was done to filter out redundant features or features with low predictive power, which may improve the prediction accuracy. Removing redundant features may also improve the interpretability of regression model, making it more suitable for manual improvements and eliminating possible competitive features. The CfsSubsetEval algorithm [11] was used as the evaluation criteria for feature selection. This algorithm evaluates the worth of subsets of the features based on the predictive ability of each feature while controlling for the possible redundancy between features, making it suitable for the goal of filtering out redundant features without losing prediction accuracy. A BestFirst search method with a search termination of 5 was used to select the best subsets from the evaluator, experiments were done with three search directions:

- Forward search, which starts with an empty subset and incrementally adds features until the predictive ability does not improve anymore for 5 subsets.
- Backward search, which starts with a subset containing all features, incrementally removing them.
- Bidirectional search, a combination of forward and backward search, performing a search from both directions at the same time.

The resulting models were again tested on their accuracy and suitability as fitness functions.

Results regression algorithms. Regression accuracies and correlation coefficients were obtained for for four regression algorithms per dimension. For each algorithm 10 holdout repetitions were ran in the Weka experimenter with 66% of the data points used for training and 33% for testing, randomized for every repetition. Table 3 shows the mean absolute error (MAE) and the correlation coefficient over all 10 runs for each algorithm and each dimension. The MAE is calculated by taking the mean of the errors (the distance of the predicted value to the actual value) of each prediction on the test-set. As a baseline the ZeroR algorithm is used.
The tested regression algorithms are hardly more accurate than ZeroR for all three dimensions. For example ZeroR has a MAE of 0.44 for P, while Linear Regression and SMOreg have a MAE of 0.43 and M5Rules a MAE of 0.44. On a scale of -1 to 1, a MAE of 0.44 is large. For A, only the M5Rules accuracy is 0.01 higher than taking the mean A rating, and for D this is the case for Linear Regression and M5Rules. So these models are not very useful to accurately predict the affect of the abstract artworks in the dataset. This implicates that either the extracted features are not very effective in capturing the affect of the images or that the dataset of affect ratings was insufficient in terms of dataset size or rating consistency, due to the limitations of using the AffectButton as a measurement device or the inherent subjectivity of the affect of abstract art. Interesting though is that significant correlations were found between the affect values and the regression models. For example the linear regression model has a correlation of 0.27 with P, from which it could be interpreted that this model can account for 27% of the differentiation in P. The correlation with A is much lower, with only the Linear Regression model having a significant correlation. For D all models have a low significant

correlation as well. Thus although the prediction accuracies of the models score hardly better than ZeroR, some models do correlate slightly with the rated affect dimensions. For the P dimension this correlation is the highest, so the used features capture the P dimension better. A reason for this could be the higher consistency in the P ratings between users.

As a secondary measure, the built models were analysed on their suitability as a fitness function in an evolutionary system. Table 4 show the variety in features for all dimensions and all models per tested algorithm. The SMOreg algorithm uses all features, but is hard to interpret, because the data mining tool used (Weka) only outputs a list of the normalized weights of each feature. The normalization of these features makes it complex to use this model as a direct measure for affect on new images, because it is not clear how each feature value is normalized. The Linear Regression algorithm uses a broad set of feature types for all three dimensions, and all feature types for the P-dimension. The resulting model is a direct formula that uses calculated weights for features to determine the affect, this formula could also directly be applied as a measure for affect in other abstract images. The M5Rules model uses slightly less features and feature types, but the model is equally easy to interpret, consisting of several formulas similar to the linear regression models, with a limited amount of rules to determine which formula to use.

Overall it can be concluded that the models built by the tested regression algorithms are not useful for measuring affect on the A dimension, bad for measuring affect on the D dimension and only slightly better for measuring affect on the P dimension. This gives reason to shift the focus of this study to the P dimension, because this is the only dimension for which a more or less usable affect measure can be determined based on these features. For the sake of completeness the other dimensions are still used in subsequent experiments, but the main focus will be on evolving abstract art for the affect dimension P. As for what regression model to use, the model built with the Linear Regression algorithm is selected. It has the highest correlation with P (and A, D), has a high variety in used features and can directly be used as an affect measure.

Table 3. Mean absolute error (MAE) and Pearson correlation coefficient (CC) for each dimension and each regression algorithm.

	Pleasure		Arousal		Dominance	
	MAE (SD)	CC (SD)	MAE (SD)	CC (SD)	MAE (SD)	CC (SD)
ZeroR	0.44 (0.01)	0.00 (0.00)	0.72 (0.01)	0.00 (0.00)	0.48 (0.01)	0.00 (0.00)
LR	0.43 (0.01)	0.27* (0.05)	0.72 (0.01)	0.05* (0.02)	0.47 (0.01)	0.11* (0.04)
SMOreg	0.43 (0.01)	0.26* (0.04)	0.72 (0.01)	0.03 (0.03)	0.48 (0.01)	0.10* (0.04)
M5Rules	0.44 (0.01)	0.25* (0.04)	0.71 (0.01)	0.04 (0.03)	0.47 (0.01)	0.12* (0.04)

*= significant difference in correlation compared to ZeroR

Table 4. Number of features and feature types used in the built regression models.

	Pleasure		Arousal		Dominance	
	Features	Types	Features	Types	Features	Types
LR	45	8	32	5	21	7
SMOreg	69	8	69	8	69	8
M5Rules	36	6	9	3	19	5

Results attribute selection algorithms. For the second experiment the CfsSubsetEval attribute selection algorithm was tested with different search directions to determine their effectiveness in improving the accuracy of the linear regression model and used as a method to filter out redundant features. The setup of this experiment was similar to the first experiment, running 10 iterations per search direction, per dimension, with a randomized 66% train/test split. The results can be found in Table 5. Using the CfsSubsetEval attribute selection algorithm hardly improves the accuracy of the

linear regression model, showing a 0.01 decrease in the MAE for all dimensions for all search directions. There is a slight increase in the correlation with P when selecting attributes through a backward search. Table 6 shows the variety in features used by the resulting regression models. When using forward or bidirectional search, the resulting model uses 7 features and 3 different feature types for P (general, color and shape measures). The backwards searched model uses 7 types for P (general, color, gray-scale, edge, shape, binary and texture), omitting only one feature type. For the A and D dimensions the backward search method also selects a larger variety of features. Although it uses less features than the non-filtered Linear Regression model, the attribute selection algorithm filters out redundant features. Looking at the final goal of using these models as fitness functions, eliminating redundant features is useful because it makes for a less complex fitness function and it reduces the possibility of the evolutionary process to optimize on one visual aspect that is represented by several features. So given that all built models had a similar prediction accuracy, the model built with a backward searched CfsSubsetEval attribute selection algorithm was selected as the most suitable affect measure for the P dimension. For completeness the models resulting from this selection algorithm were also used as the measures for A and D.

Table 5. Mean absolute error (MAE) and Pearson correlation coefficient (CC) of the linear regression models for each dimension with or without CfsSubsetEval attribute selection in three search directions.

	Pleasure		Arousal		Dominance	
	MAE (SD)	CC (SD)	MAE (SD)	CC (SD)	MAE (SD)	CC (SD)
No selection	0.44 (0.01)	0.00 (0.00)	0.72 (0.01)	0.00 (0.00)	0.48 (0.01)	0.00 (0.00)
Forward	0.43 (0.01)	0.24 (0.04)	0.71 (0.01)	0.06 (0.03)	0.47 (0.00)	0.14 (0.04)
Backward	0.43 (0.01)	0.27 (0.04)	0.71 (0.01)	0.05 (0.03)	0.47 (0.01)	0.12 (0.04)
Bidirectional	0.43 (0.01)	0.24 (0.04)	0.71 (0.01)	0.06 (0.03)	0.47 (0.01)	0.14 (0.04)

Table 6 Number of features and feature types used in the built regression models for each dimension with or without CfsSubsetEval attribute selection in three search directions.

	Pleasure		Arousal		Dominance	
	Features	Types	Features	Types	Features	Types
No selection	45	8	32	5	21	7
Forward	7	3	7	3	5	3
Backward	29	7	27	5	14	5
Bidirectional	7	3	7	3	5	3

Interpreting the learned affect measures. Table 7 shows the features that will be used in the affect measures for each dimension. The weights in these table are based on normalized feature values, so their absolute weight can be compared. At the top are the features that have the highest positive impact on the affect dimension, at the bottom the features with the highest negative impact. The found importance of colour and shape features is similar to the findings of the studies in table 1. It is interesting to look closely at these features because the combination of all these features will define which generated images will be measured as having a high or a low fitness. So it gives an indication of what visual properties will be favoured in the evolutionary process. The feature with the largest absolute weight is the mean brightness, higher brightness implicates higher pleasure and vice versa, not a surprising relation. But take for example the amount of shape-type features used as a measure for P, especially the features based on the mean properties of all found shapes have a large weight. For example, the mean area of all shapes has a negative impact on P, while a higher standard deviation of this mean is positively related with P, as is the mean perimeter of the found shapes. This indicates that images with a high variation in shape area, that on average have a small area but a large perimeter, are

Table 7. Features used in the learned affect measures and their normalized weight. SD = standard deviation.

Pleasure		Arousal		Dominance	
Feature	Weight	Feature	Weight	Feature	Weight
Brightness Mean	2.3517	Shapes roundness SD	1.7164	Saturation Mean	0.4348
Shapes area SD	2.0887	Shapes area Mean	1.6058	Edge gray value SD	0.3909
Shapes perimeter Mean	0.6873	Mean gray value	0.9383	Green Mean	0.3483
Green Mean	0.5479	Main shape circularity	0.77	Shapes compactness Mean	0.342
Shapes aspect ratio Mean	0.4593	Shapes angle SD	0.7209	Brightness SD	0.281
Shapes Form factor Mean	0.4281	GLCM Angular Second Moment	0.5532	Saturation SD	0.2542
Percentage hot colours	0.3399	Shapes aspect ratio SD	0.4804	Blue Mean	0.1946
Maximum gray value	0.2866	Saturation Mean	0.4493	Main shape area	0.1396
Red SD	0.269	Main shape distance to image center	0.285	Saturation Mode	-0.1051
Edge gray value SD	0.2437	GLCM Inverse Difference Moment	0.2619	Main shape y-position	-0.1293
Laplacian gray value SD	0.2347	Saturation SD	0.233	GLCM Inverse Difference Moment	-0.1832
Main shape area	0.2157	Blue Mode	0.1962	Blue SD	-0.1847
Number of different colours	0.2035	GLCM Entropy	0.1694	Shapes roundness M	-0.4885
Main shape solidity	0.1321	Number of different colours	0.1676	Laplacian gray value SD	-0.5563
Hue Mode	0.1162	Red Mode	0.148		
Binary gray value SD	0.1016	Orange, blue or green as main colour	0.1467		
Feret angle	-0.1807	Main shape solidity	0.1293		
Main shape convex hull	-0.2037	Green Mode	-0.114		
Main shape short side	-0.2184	Feret angle	-0.1614		
Main shape y-position	-0.2673	Red, orange, blue or green as main colour	-0.1616		
Saturation Mean	-0.3029	Shapes count	-0.1633		
Edge gray value Mean	-0.3463	Brightness Mode	-0.2262		
GLCM Angular Second Moment	-0.3591	Green Mean	-0.593		
Minimum gray value	-0.3748	Laplacian maximum gray value	-0.6715		
Shapes roundness SD	-0.4544	Laplacian minimum gray value	-0.7081		
Brightness SD	-0.4987	Shapes aspect ratio Mean	-0.9632		
Laplacian maximum gray value	-0.5335	Shapes area SD	-1.2483		
Blue Mean	-0.7593	Shapes compactness SD	-1.5994		
Red Mean	-1.7186				
Shapes area Mean	-2.0677				

measured as having a high P, thus will be rated as fitter when optimizing for a high P. Such interpretations will give an idea of what type of images will be preferred when using these measures in to evolve art. However, caution should be taken in interpreting these findings as actual reliable measures for the affect of abstract art in general. Although these models are the best measures that could be found on this dataset, with these specific features, the accuracy of the original models in predicting affect is still very low. The use of the mean green value as a measure of P is an example of why this caution is necessary. It could be that images rated high on P actually have brighter green colours, but it could also be that the images rated low on P have very little green colours. But the model still uses this feature as an important measure because predicting that images with at least some green colour do not have a low P will already increase the accuracy. Another important notion is that the weights of these features are based on the range of the feature-values as found in the images in the dataset. So where the representation and rating of affect on the PAD dimensions is limited to -1 and 1, these measures are not limited to this range and specific combinations of features could result in measured affect values outside of this range.

3.6. Conclusion on learning affect measures

Affect measures for the PAD dimensions were learned from the corpus of affective abstract artworks. Due to the low correlations and low prediction accuracy of the regression models that these measures are based on, they are not expected to be very effective measures for affect, especially not on the A and D dimensions. The measures can however directly be applied as fitness functions in an evolutionary system and use a wide variety of features to measure affect, which accommodates for a broad search space for the evolutionary system.

4. Evolving affective abstract art

The learned affect measures were applied as fitness function within an evolutionary system that generates artwork in a restricted style. This section first describes the setup of the evolutionary system and the way the affect measures are integrated in the evolutionary process as fitness functions. It is followed by experiments on optimizing for different affect goals and experiments into manually adjusting the measures and evolutionary parameters. The section concludes with the setup and results of a validation web survey, used to determine the effectiveness of the evolutionary system in generating affective art.

4.1. Evolutionary system

A novel system to generate art through an evolutionary algorithm was developed. Goal of this was to explore the possibilities of generating art with different affective properties within one simplistic style of art. This style was designed to look more natural than the common generative art based on mathematical functions and principles.

Genotype and phenotype. The system generates images by drawing a set amount of layers in a sequential order on the same canvas. Seven layer types were designed, which can be found in figure 8, each type draws a different type of shape(s) based on 8 values. For this a list of floating points, with a precision of eight, is used (the genotype). The amount of genes is dependent on the amount of layers, with two genes encoding for the canvas and eight genes per layer that encode for its appearance, as seen in figure 9. Each layer has one shape-type value, four values to determine its colour and three form values to determine the size, position and complexity of the shape. The shape-type value is used to determine which of the seven shapes to draw, including an eight type that turns the layer off (nothing will be drawn on that layer). The colour values are used to determine the hue, saturation,

brightness and opacity of the drawn shape. The position, rotation, size and form of the shape are dependent of the shape-type and the three form values. As an addition a randomized list of integers ranging from 1 to 1000 is used, this was done to achieve a perceived sense of randomness while still keeping the system deterministic. The form values are used to determine the properties of the shape, either directly, in combination with each other, or as a pointer to a value in the integer list. This setup makes it possible to draw a vast arrange of different forms of one shape based on only three floating point values. Finally, the two canvas values are used to determine the aspect ratio of the canvas (since this image feature is also used in the learned measure for P) and the brightness of the canvas. The maximum side length (either width or height depending on the aspect ratio) of the canvas is set to a fixed 500 pixels and the minimum to 250 pixels.

Figure 8. Examples of all shape types and the resulting combination of these layers on the right.

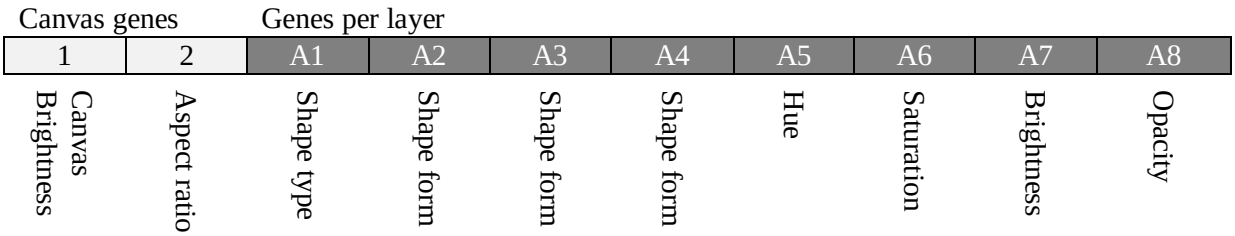

Figure 9. Chromosome structure for an individual.

Evolutionary setup. The system uses a uniform crossover method and point mutation method. Parents are selected through tournament selection, initialization is done through a random generation of a floating point value between 0 and 1 for each gene. As a fitness function (combinations of) the found affect measures for PAD were used. After evolving (or initializing) a new generation, all individuals were saved as images in .tif format, from which the necessary image features were extracted. For this feature extraction the exact same method was used as in extracting the features from the original dataset of artworks. The extracted feature values were then used to compute the PAD ratings. Each individual was then assigned its fitness based on a set affect goal. This goal could be any value for each PAD dimension, either on multiple dimensions or a single dimension. For example if the goal was set to optimize P to as high as possible and not using the A and D dimensions, the individual with the highest P rating would be the fittest. But more precise goals, like optimizing P to 0.2, A to -1.0 and D to 0.5 were also possible, giving great flexibility in affect goals. Last part of the setup was determining which shape types could be drawn and how many layers each individual would have. Because the number of layers influences the number of genes and the used shape-types influence the possible feature values of the individuals, these were both set to a fixed value for all experiments. But there is a shape-type that turns a layer off, so although there is a set number of layers, there is still the possibility that artworks display less layers.

Experimental setup. Preliminary tests were run to determine the evolutionary parameters (population size, crossover rate, mutation rate and tournament size), these parameters were kept consistent throughout all experiments and can be found in table 8. First, an experiment was done to determine the affect range of the images generated by the system. After that a set of experiments was conducted with different affect goals, optimizing independently for affect as high as possible on all dimensions, a set

1.0 value for P and a set -1.0 value for P. Due to the poor accuracy of the learned affect measures, the choice was made to only optimize for the outer boundaries of the affect dimensions. The set values were used as optimization goals to supress feature exploitation that resulted in measured P values far outside the -1 to 1 scale. Additional runs were done with experimental affect goals and evolutionary parameters, including optimizing for multiple dimensions, restricting the amount of layers in the phenotype and optimizing without controlling for colour. To validate the actual affect of the output, a second web survey was used to gather affect ratings on a sample set of evolved artworks.

Table 8. Evolutionary parameters used for all experiments.

System Parameters	
Representation	List of floating points
Initialization	Random per gene
Survivor selection	None
Parent selection	Tournament
Recombination	Uniform crossover
Mutation	Point mutation
Numeric parameters	
Population size	100
Tournament size	5
Crossover rate	0.5
Mutation rate	0.1
Individual parameters	
Layer amount	6
Shape types	8
Genes	50
Maximum canvas side	500 pixels
Minimum canvas side	250 pixels

4.2. Experiments in evolving for affect

Experiment 1: Affect range of an initialization generation. A limitation of using predefined shapes is the restricted range of possible outputs. A consequence of this is that these restricted outputs may not cater for the whole affect range. For this purpose the first experiment was conducted to determine the measured affect over 1000 individuals of an initialization generation. Examples of these first generation artworks can be found in figure 10, a histogram of the measured affect can be found in figure 11.

As the histogram shows, the 1000 individuals have a broad range of affect on the P dimension. The individuals have a mean measured P of 0.019 with a standard deviation of 0.625. The A measure (M= -0.396, SD = 0.582) has a bias towards negative affect and D (M = -0.120, SD = 0.175) has a very small measured range, with almost all artworks measured between -0.5 and 0.2. Individuals were measured as having an affect outside the original range of -1 to 1 on the P and A dimensions, with affect as high as 3.43 for P and as low as -7.02 for A. This means that the output of the evolutionary system has visual properties that were not present in artworks in the original dataset, which results in feature values in a broader range than in the original dataset. These higher or lower feature values were not accounted for in the models that the affect measures are based on. Caution should thus be taken when optimizing for high P or low A, because these features could be exploited to reach very high or very low affect, as can be seen in experiment 2.

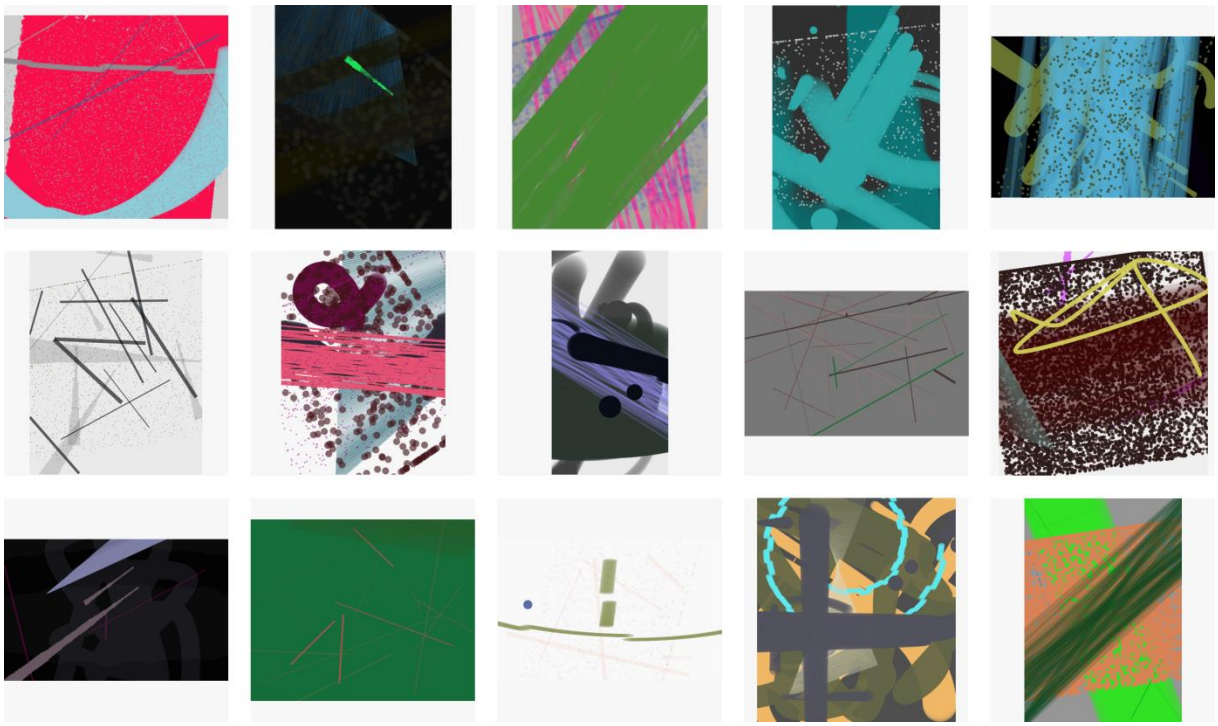

Figure 10. 15 individuals from an initialization generation.

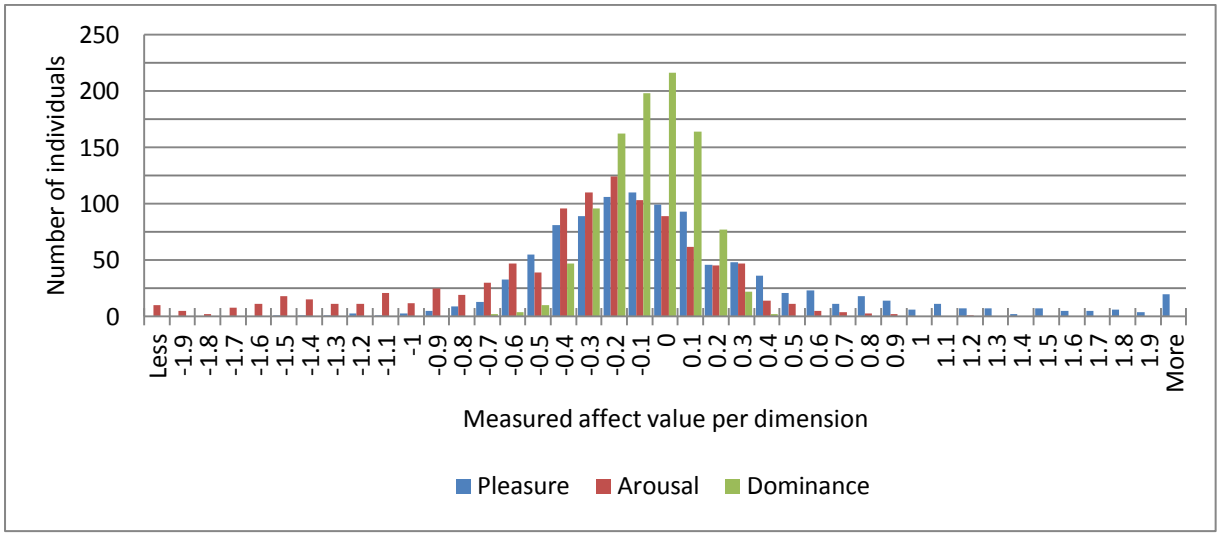

Figure 11. Histogram of the measured affect on the PAD dimensions for 1000 initialization individuals.

Experiment 2: High positive affect. A second set of experiments was conducted for three different affect goals: high P, high A and high D. The fitness for individuals in these experiments was calculated by directly taking the measured affect on the goal dimension, with no upper or lower limit for the fitness value and ignoring the other dimensions. Ten runs were done per goal, with a generation limit of 30 generations. Figure 12 shows the mean measured affect over 30 generations and 10 runs for the three dimensions when optimizing for high P. Similar results where reached when optimizing for the other two dimensions. With the mean affect for A optimizing from -0.380 in the first generation to 1.82 in the last generation. For D this optimization went from -0.125 to 0.680. These results show that the evolutionary setup is indeed capable of optimizing for a measured high affect on the three dimensions independently. Looking at high P as the affect goal, the individuals optimize for values outside the original scale of -1 to 1, reaching a mean P value of over 7.0. There is also a negative impact on A when optimizing for high P. Since a positive correlation was found between P and A in

the ratings of the original dataset, this may be another indication that these affect measures are poor for measuring actual affect or that the restricted style of the evolved artworks does not cater well for these affect measures. Since the measures for A and D were determined to be poor, the results on these dimensions will not be discussed in depth.

Figure 13 shows the fittest individual of each run after 30 generations. There is a lot of similarity in these images; bright green colours, diagonal lines and at least some black colour (either in shapes or the canvas) are preferred as visual properties. Comparing the visual properties of these fittest individuals with the visual properties of the artworks in the original dataset that were rated high on P, there is a clear similarity in the overall brightness. The bright green colour and the large shapes however are not seen in the original dataset. As explained in section 3.5 this might be a result of the poor accuracy of the learned models, but it might also be a result of the exploitation of certain features. To study this exploitation, the feature values and the measured P of all individuals in one evolutionary run were analysed. This particular evolutionary run resulted in a fittest individual with a P of 14.14. The analysis showed a clear exploitation of one feature, the mean perimeter of all found shapes. For the fittest individual this feature value contributed a weighted 13.71 to the total measured P, all other image features thus only accounted for 0.43 of the measured P. Analysis of other runs confirmed that the overfitting was mainly caused by this feature. Because the high values on this feature are not present in the original dataset, but a consequence of the method of learning affect measures and using them as fitness functions, a third experiment was conducted in which this overfitting is suppressed.

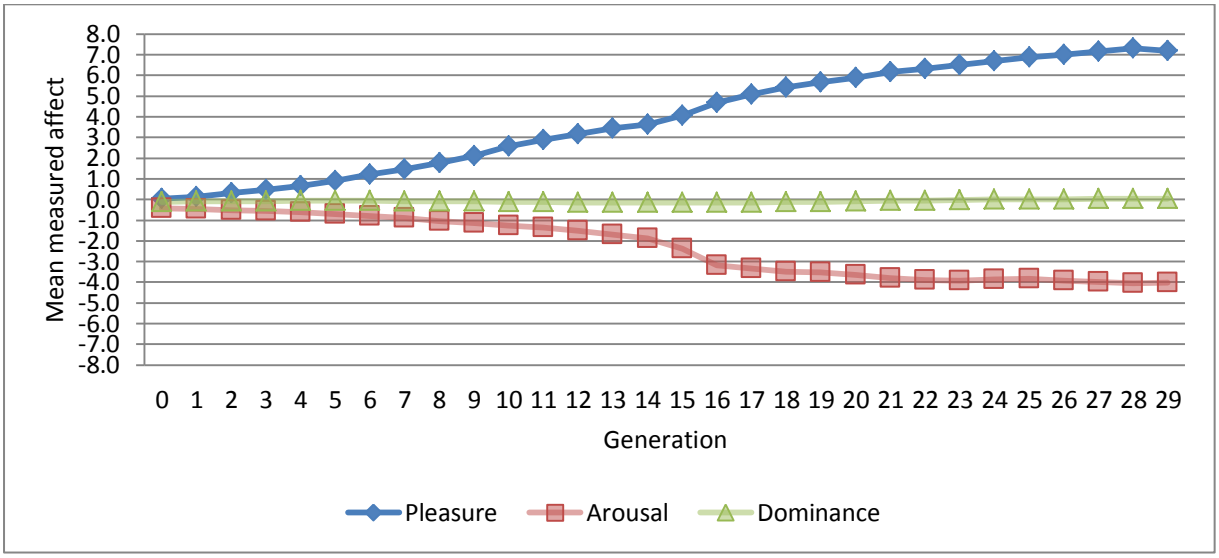

Figure 12. Mean measured affect over 10 runs optimizing for high P.

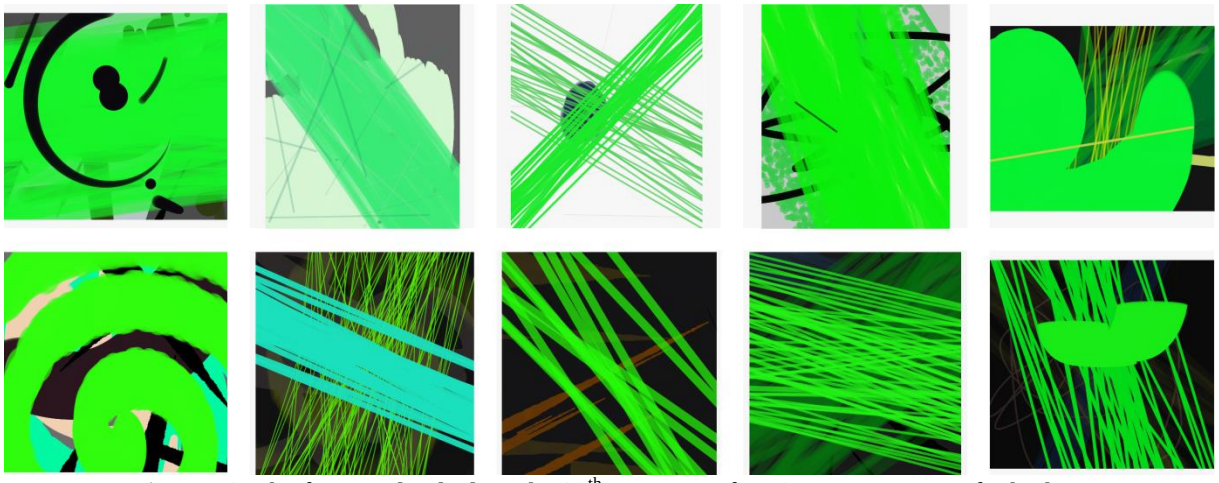

Figure 13. The fittest individuals in the 30th generation for 10 runs optimizing for high P.

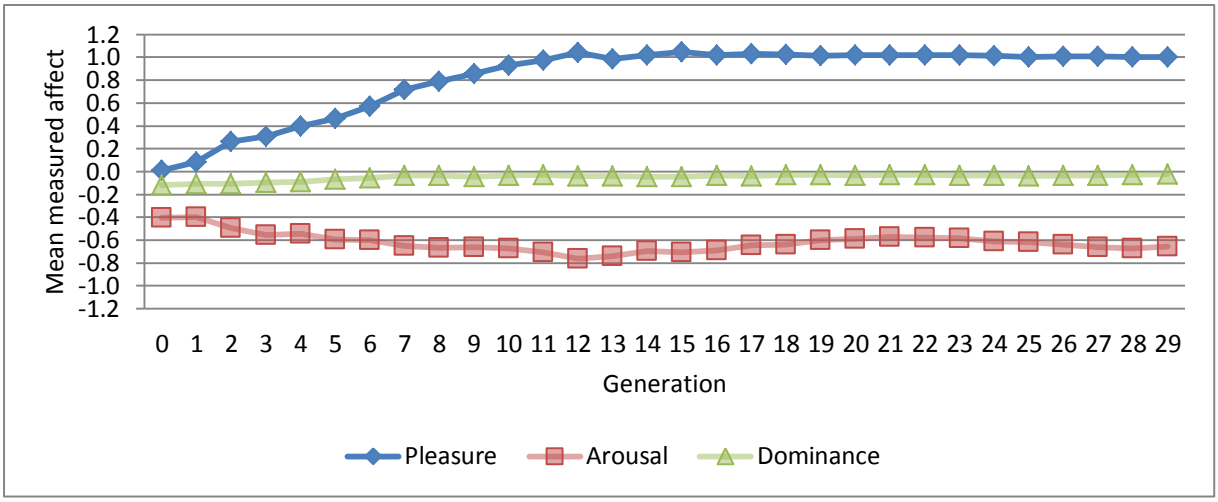

Figure 14. Mean measured affect over 10 runs optimizing for P 1.0.

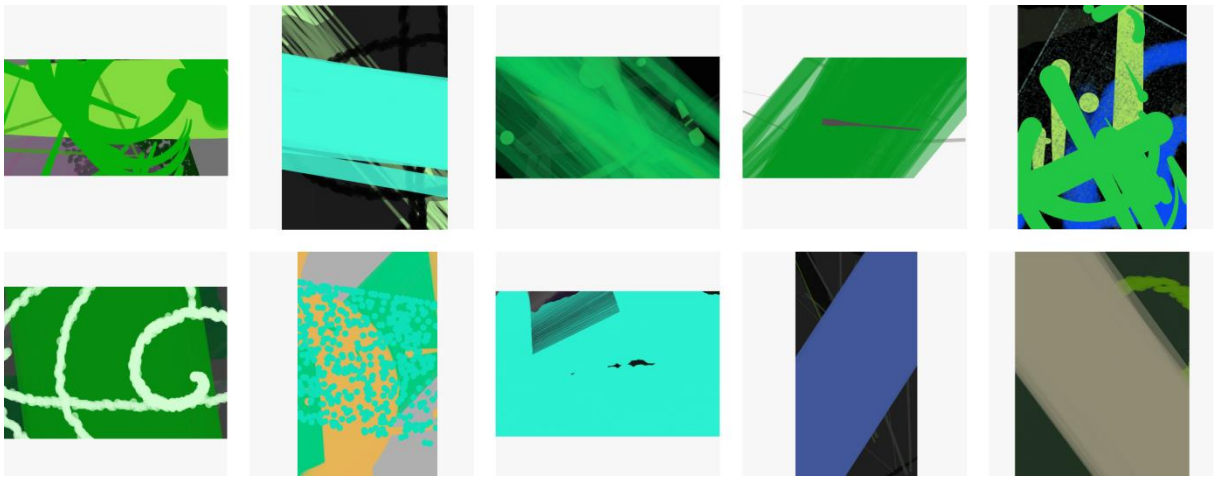

Figure 15. The fittest individuals in the 30th generation for 10 runs optimizing for a P value of 1.0.

Experiment 3: Optimizing for P 1.0. To suppress overfitting on the P dimension the optimization goal was set to a P value of 1.0. For this, the following formula was used to calculate the fitness based on the measured P value:

$$F(x) = 1 - |1 - P(x)| \quad (1)$$

So an individual with P 1.3 gets assigned the same fitness as an individual with P 0.7, thus having the same chance to be selected as a parent for the next generation. Instead of simply assigning a fitness of 1.0 to individuals with a P higher than 1, this method actually oppresses the exploitation of features, so that other features may have a bigger influence on the fitness. This may lead to a higher variety in the fittest individuals without altering the learned affect measure.

Again ten runs were done, optimizing for a P of 1.0 with a generation limit of 30. Figure 14 shows the optimization process and figure 15 shows the resulting fittest individuals. The process shows that P indeed optimizes towards 1.0 and is corrected when going over this value. The generated images are more varied than when optimizing for high P. The green colours, diagonal lines and large shapes are still present, but there is more variation between them. Interesting is the difference in the aspect ratio of the canvasses. In the previous experiment the aspect ratio was close to 1 (equal width and height), while half of these canvasses have an aspect ratio of 0.5 or 2.0, which might be due to the fact that a bigger canvas size caters for larger perimeters of shapes. Comparing the images with the original

dataset they show more resemblance with the artworks rated high on P, mostly due to the higher variation in color. But the two artworks in the bottom right corner of figure 15, which both have a measured P of 1.0, are more similar to the original artworks rated low on P than to the original artworks rated high on P. So while supressing feature exploitation results in a higher variety of artworks, it also increases the chance of generating artworks that do not look like they have a high positive affect. It is hard to determine whether this problem is due to the poor affective measure, or that it is a consequence of the restricted style of the evolutionary system.

Experiment 4: Optimizing for P -1.0. A fourth experiment was conducted to evolve artworks with negative P. Because preliminary tests showed that feature exploitation influenced the generated images, resulting in empty black and gray artworks, the affect goal was set at P -1.0. Ten runs were done with a generation limit of 20. Figure 16 shows the resulting 10 fittest individuals, all having a measured P close to -1.0. Noticeable is that these artworks are generally darker and have more small and detailed shapes, while still maintaining some variation in visual properties. The dark colour is something that is also present in the low P rated artworks in the original dataset, the smaller shapes are present in some of these original artworks. Overall the artworks resemble the original dataset better than when optimizing for P 1.0.

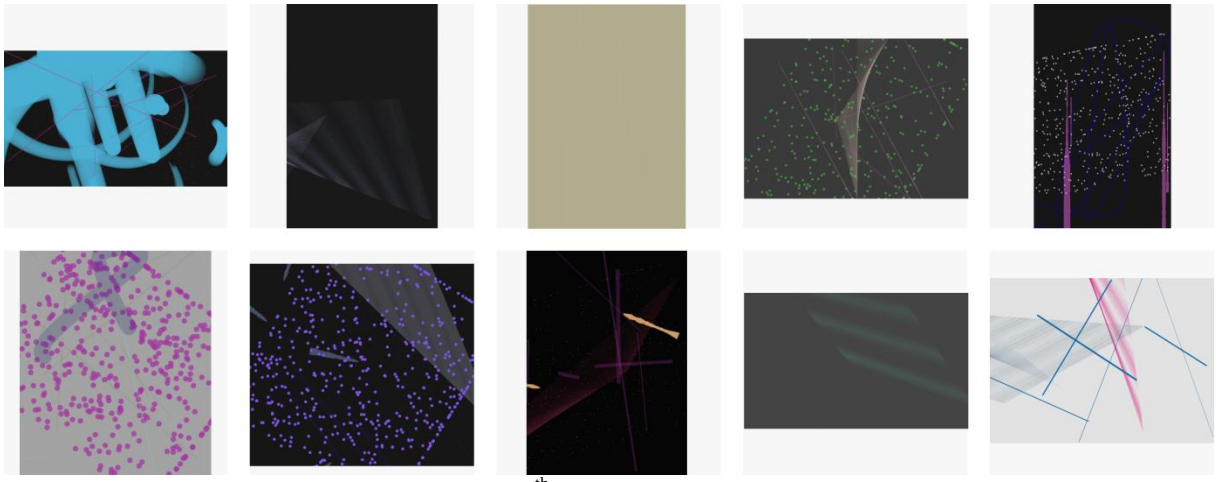

Figure 16. The fittest individuals in the 20th generation for 10 runs optimizing for a P value of -1.0.

Experiment 5: Multidimensional affect goals. In the last formal experiment on generating affective art, images were evolved by optimizing for an affect goal on all three dimensions. The reason for this is two-fold. The first reason is to evolve more interesting artworks. For example, optimizing for P 1.0 results in individuals with low A (as can be seen in figure 12). This means that features that have a positive weight for P either have a negative weight for A or influence other features in such a way that it negatively impacts A. Optimizing for a value of 1.0 on both of these dimensions would mean that these features compete with themselves, so other features get more important for the fitness of the individuals. This might result in more interesting and more varied output. A second reason for these multidimensional goals is the positive correlation between P and A, and P and D in the user ratings. This implicates that a higher perceived affect on the A and D dimensions might aid in optimizing for a higher perceived affect on the P dimension. Although this correlation is not directly apparent in the affect measures (that were learned independently of each other), it is interesting to see if the A and D measures can be used to improve the perceived P. For this set of experiments two different multidimensional affect goals were used, one goal was optimizing for P 1.0, A 1.0 and D 1.0, the other was the opposite, optimizing for P -1.0, A -1.0 and D -1.0. A weighted mean was used to calculate the fitness for these runs, the fitness for PAD 1.0 was calculated as follows:

$$F(x) = \frac{3 - |1 - P(x)| - |1 - A(x)| - |1 - D(x)|}{3} \quad (2)$$

And the fitness when optimizing for PAD -1.0 was calculated using:

$$F(x) = \frac{3 - |-1 - P(x)| - |-1 - A(x)| - |-1 - D(x)|}{3} \quad (3)$$

For each goal ten runs were done, with a generation limit of 20. The mean measured affect over the evolutionary runs is seen in figure 17 and 18. The last figure shows that the multidimensional goals hardly compete when optimizing for a value of -1.0, with all three dimensions reaching a mean fitness close to -1.0. Figure 18 shows that A and D do not optimize very well towards 1.0 when using all three dimensions as measures compared to optimizing them separately to as high as possible. The P dimension still approaches the goal value of 1.0, but takes more generations to do so. But the fact that this happens while A and D also reach positive affect values, means that different features are preferred in the evolutionary process than when optimizing only for P 1.0. This however does not lead to greater variation in the output, as can be seen in the examples in figure 19. There is great consistency in the shapes and the used colours, which could be explained by the fact that there are simply less (combinations of) features that positively influence all three affect dimensions. The bottom row of figure 19 shows a similar consistency in output, resulting in mostly dark and empty images.

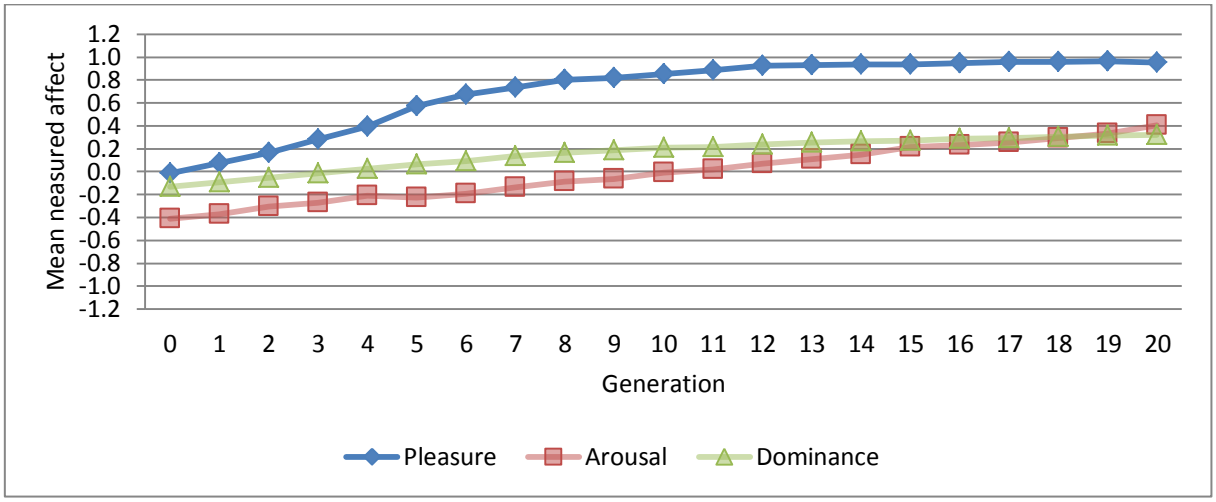

Figure 17. Mean measured affect over 10 runs optimizing for P, A and D to value 1.0.

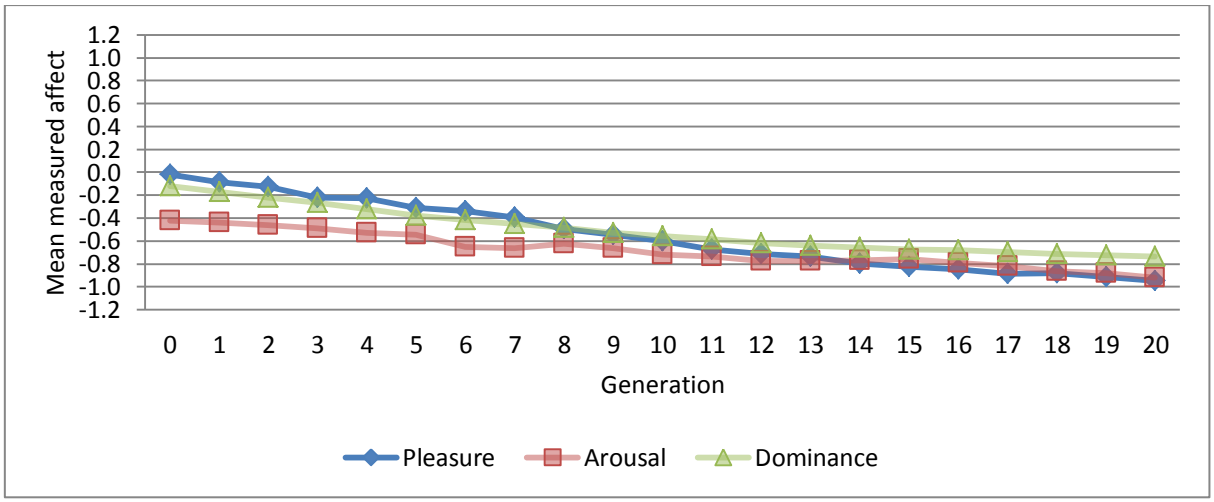

Figure 18. Mean measured affect over 10 runs optimizing for P, A and D to value -1.0.

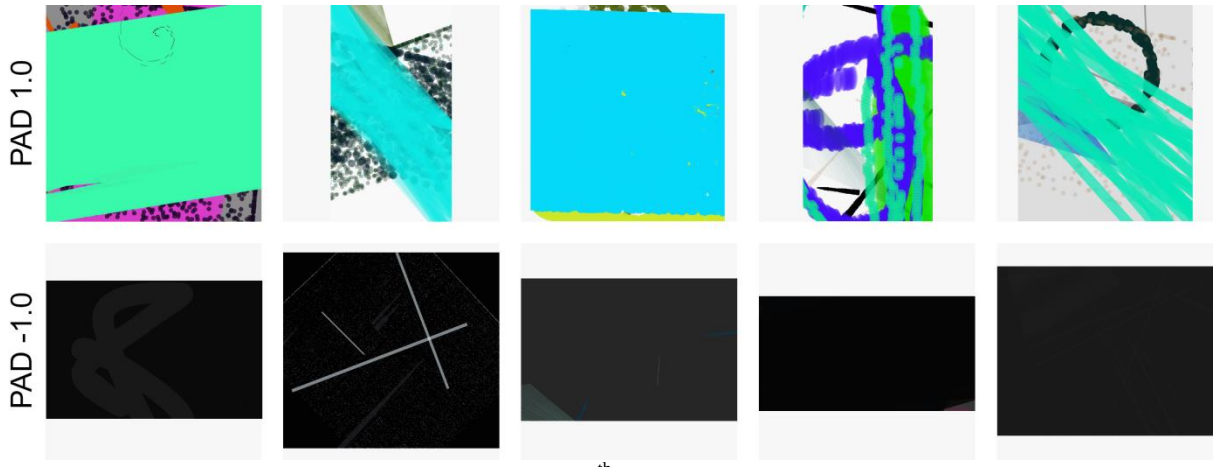

Figure 19. Examples of the fittest individuals in the 20th generation when optimizing for a value 1.0 and -1.0 on all affect dimensions.

4.3. Experiments in manually adjusting the evolutionary setup

Next to the formal experiments on evolving for different affect goals, two sets of experiments were done to explore the possibilities of manually adjusting the fitness function (and thus the evolutionary process) and tweaking the amount of layers. These experiments and the resulting artworks are shortly discussed.

No colour. Because the affect measures cause a very distinct preference in colour during the evolutionary process, experiments were done where this process was disrupted. Three adjustments were made:

1) The weights for the colour features in the affect measures were set to 0.
2) The weight for the colour features and the gray-scale features in the affect measures were set to 0.
3) The genes that encode for the saturation of layers were not expressed in the phenotype, resulting in gray-scale images.

It should be mentioned that these adjustments influence the range of possible measured affect and may thus influence the evolutionary process in other ways than just neglecting colour. For each adjustment 5 runs were done, optimizing for a high P, with a generation limit of 20.
Examples of the fittest individuals can be seen in Figure 20. The consequences of the adjustments are already apparent in the measured affect of the initialisation generation. With a mean P of -0.786 for adjustment 1, 0.513 for adjustment 2 and 0.281 for adjustment 3 for all individuals in the initialisation generations. This mean P was 0.019 when no adjustments were made. Still, the adjusted runs optimized well, with all resulting fittest individuals reaching P values over 1.0. The most interesting and varied output came from adjustment 2, without controlling for colour and gray-scale the preference for lots of bright colours is gone. The result is artworks that have more natural and varied colours compared with the original and adjustment 1 runs, where the clear preference for brightness makes for a more homogeneous set of generated artworks. The generated artworks without saturation in the phenotype display the same visual properties (like diagonal lines) as the unadjusted runs. So the results of adjustment 2 show that manual adjustment of the fitness function can improve the variation and the interestingness of the artworks, while still optimizing for high P.

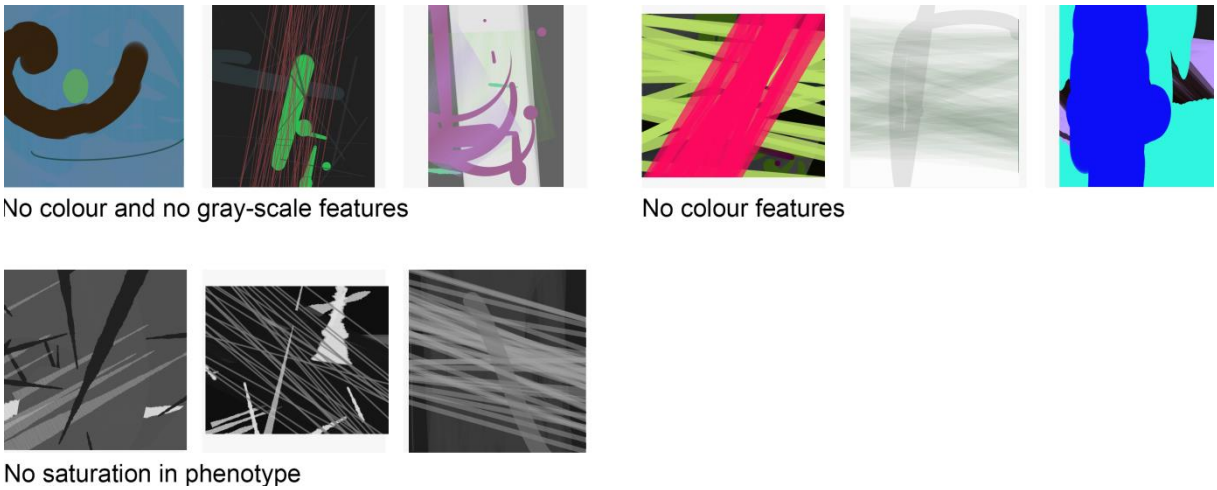

Figure 20. Examples of the fittest individuals in the 20th generation when optimizing for a P value as high as possible with colour features removed from the fitness function or no saturation in the phenotype.

Maximum amount of layers. A second variation on the original evolutionary setup was adjusting the maximum amount of layers drawn. The maximum was set to 1, 2, 3, 12, 24 or 48 layers, doing 5 runs for each setting. The other parameters were kept identical to table 8. The runs optimized for P, A and D 1.0, with a generation limit of 20. Figure 21 shows examples of the resulting artworks. Unsurprisingly there is a lot more happening in the artworks with more layers, but the large shapes with bright green and blue colours, that were also seen in experiment 5, are dominant in all settings.
For the initialisation individuals the amount of layers influences the measured P, A, D values. Using only 1 layer the mean PAD for the initialisation generation is -0.689, increasing to 0.446 and 0.305 for 2 and 3 layers. For 12, 24 and 48 layers these means are 0.100, 0.087 and 0.0823. In the 20th generation this initial difference in measured PAD is gone, with all runs optimizing to a mean PAD of approximately 0.5. So even though there is a lot more going on in background of the individuals with more layers, this one shape is enough to optimize for PAD values of 1.0. This is might be due to top layers masking the lower layers, so whether an individual has 3 layers or 48, as long as the last drawn layers have a large area and high opacity they become the main visual property. But while the actual output may look similar, individuals with more layers also have more genes, thus their evolutionary process differs. A possible adjustment to generate more varied art might be supressing the appearance of this shape or giving the drawn layers a decreasing opacity value, so that each layer has the same chance to influence the measured affect.

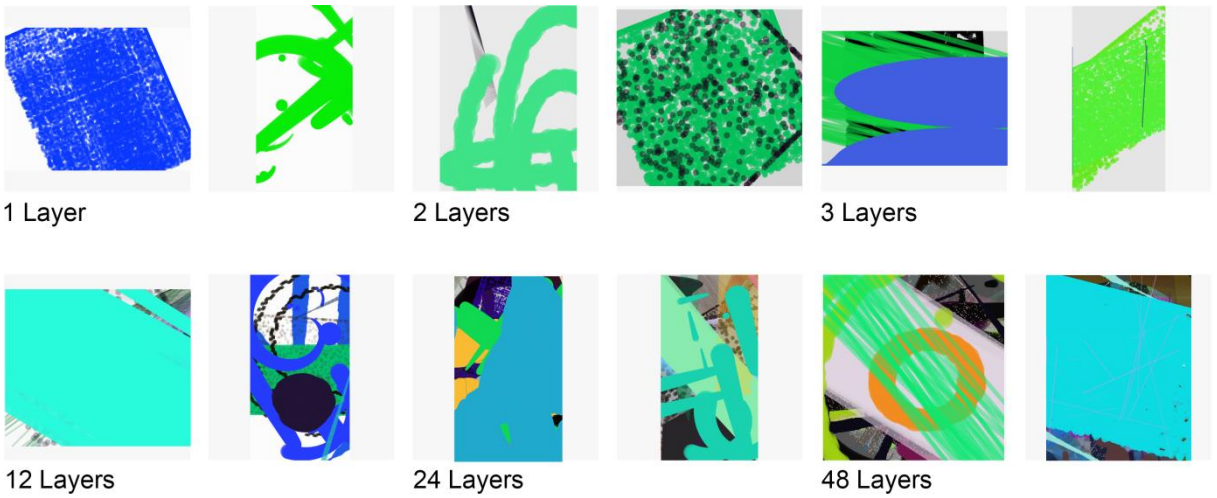

Figure 21. Examples of the fittest individuals in the 20th generation when optimizing for PAD values of 1.0 with a set maximum number of layers.

4.4. Conclusion on experiments in evolving affective art

With the introduced evolutionary system, several artworks were generated based on the affect measures. The affect measures were used to optimize for different affect goals, with a focus on the P dimension. A varied range of artworks was seen in the individuals of the initialisation generation, showing that the combination of the phenotype and the affect measures could generate images on a broad range of measured affect. Overfitting when optimizing for high P resulted in very similar artworks, optimizing to a set value of 1.0 or -1.0 on P supressed this overfitting and resulted in more varied artworks, but possibly increased the chance of evolving artworks that are not perceived as 1.0 P or -1.0 P. Artworks generated when optimizing for multidimensional affect goals showed less variety. Additional experiments showed the opportunities of adjusting the affect measures or evolutionary parameters to generate more interesting and varied art. So although the used affect measures were poor measures for actual affect, they can be used as fitness functions, either combined or independently, to evolve abstract artwork.

4.5. Validation web survey

To test whether the evolved abstract artworks are actually perceived as having the affect for which they were optimized, a second online web survey was conducted with a selection of generated artworks. These were rated by participants on their affect and these ratings were compared with the intended affect for which the artworks were optimized. This determines whether the system was effective in generating art with certain affect goals.

Setup. A selection of 90 unique artworks evolved for different affect goals were used for this survey:

- 30 artworks from an initialisation generation, generated specifically for this survey. To test whether the limited style of the evolutionary system caters for the whole range of perceived affect on the three dimensions.
- 20 artworks optimized for P 1.0. To test whether the affect measure for P is an effective fitness function for evolving artworks that are actually perceived as having high P. These artworks were chosen instead of the ones optimized for high P because they showed more variation.
- 20 artworks optimized for P -1.0. To test whether the affect measure for P is effective for evolving artworks perceived as having low P.
- 10 artworks optimized for PAD 1.0. To test whether the additional optimization goals of A 1.0 and D 1.0 help in evolving artworks that are perceived as having high P.
- 10 artworks optimized for PAD -1.0. To test whether these additional goal also help in evolving artworks that are perceived as having low P.

For this survey the exact same setup was used as for the first web survey, participants were asked to rate 25 generated abstract artworks using the AffectButton.

Results. A total of 37 participants were recruited online, of which several also participated in the first survey. Participants ranged in age between 21 and 78 (M = 36.32, SD = 15.36; 25 men, 12 women). 925 artworks were rated, with each artwork receiving 9, 10 or 11 unique ratings. Figure 22 shows a histogram of the ratings for the artworks in the initialisation group, again an additional 3D-visualisation using the mean ratings per artwork was used to gain insight into the combined affect ratings. The spread of affect ratings resembles the spread for the original dataset (Figure 5 and 6). The generated artworks were generally rated to have more negative affect, with the biggest difference that less artwork were rated as having P 1.0 or A 1.0. Comparing the mean ratings per artwork of the original artworks and the initialisation generation shows a similar spread over the combined

dimensions, with the generated images rated slightly more negative, especially on A. Correlations were found in the ratings for P and A ($r(310) = 0.155$, $p < 0.01$) and P and D ($r(310) = 0.307$, $p < 0.001$), which were also found in the ratings for the original artworks. The ratings on this small set of initialisation generation artworks show that the evolutionary system is capable of generating images for a broad range of perceived affect, despite the restricted art style. A histogram of all given P ratings for the groups optimized for P 1.0 and P -1.0 can be found in figure 23. Although slightly more artworks from the P 1.0 group were actually rated at P 1.0, the spread of ratings over the whole P range is similar for both groups. An independent samples *t*-test was conducted to test whether the P 1.0 group was rated higher on P than the P -1.0 group. No significant difference ($t_{(404)} = 0.326$, $p > 0.05$) was found between the P 1.0 ($M = 0.022$, $SD = 0.508$) and P -1.0 ($M = 0.003$, $SD = 0.452$) groups, which suggests that participants rated both groups as having similar P. So the evolutionary system did not succeed in generating images with the affect goals -1.0 or 1.0 on the P-dimension. Figure 24 shows the histograms of the ratings for the groups optimized on PAD. Again an independent samples *t*-test was conducted, which showed a significant difference of the P rating ($t_{(103)} = 6.03$, $p < 0.001$) between the PAD 1.0 ($M = 0.235$, $SD = 0.535$) and PAD -1.0 ($M = -0.190$, $SD = 0.471$) groups. There was also a significant difference found in the ratings on D ($t_{(204)} = 4.817$, $p < 0.05$), suggesting that also D was rated closer to its intended affect for both PAD 1.0 ($M = 0.122$, $SD = 0.523$) and PAD -1.0 ($M = -0.247$, $SD = 0.573$). These results confirm the hypothesis that additional optimization on the A and D dimensions helps in successfully evolving artwork with an affect goal of both low and high P.

Another way to interpret the ratings is by looking at the mean absolute error (MAE), which in this case is the absolute difference between the rated affect and the affect as measured by the affect measures. This is useful to gain a secondary insight on the effectiveness of the affect measures. Taking all rated artworks into account, ratings on P had a MAE of 0.818 ($SD = 0.530$), with the largest MAE for the P 1.0 and P -1.0 groups, respectively 0.981 ($SD = 0.531$) and 0.965 ($SD = 0.501$). Ratings on A had an MAE of 0.760 ($SD = 0.531$) for all artworks, and ratings on D had an MAE of 0.463 ($SD = 0.351$) for all artworks. Overall the lowest MAE's were apparent in the initialisation group. This can be explained by the fact that all artworks were generally rated closer to 0.0 on all dimensions. Because the artworks in the initialisation group were not optimized towards an extremity there were more artworks that were measured as having affect more closely to 0.0, resulting in an overall lower MAE. As was to be expected, these MAE's confirm the low accuracy of the affect measures.

Conclusion. Overall it can be concluded that the evolutionary system is capable of generating artwork on a broad range of affect, with a slight bias towards negative affect. The used affect measures were again confirmed to be poor, and the measures for P cannot be used as fitness functions to effectively evolve affective artworks on that dimension. However, because of the correlation in rated affect between P and A, and P and D, the affect measures for P, A and D can be used as a combined fitness functions to evolve artworks with the goal of having low P or high P, but still not with great precision.

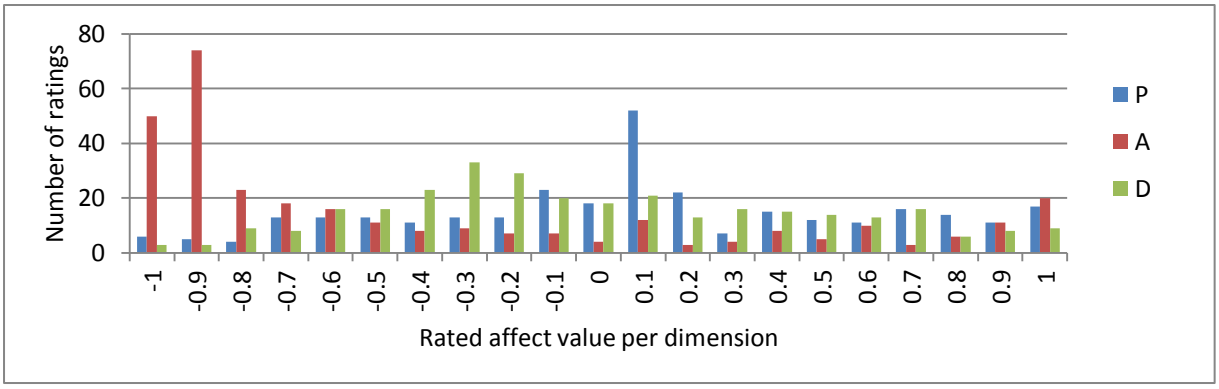

Figure 22. Histograms of affect ratings per dimension for the initialisation group.

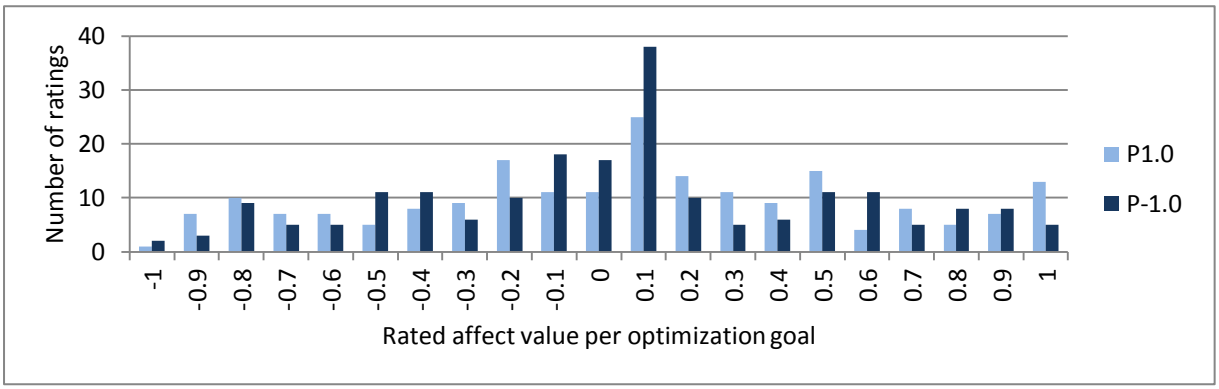

Figure 23. Histograms of affect ratings on the P dimension for 20 artworks optimized for P 1.0 and 20 artworks optimized for P -1.0.

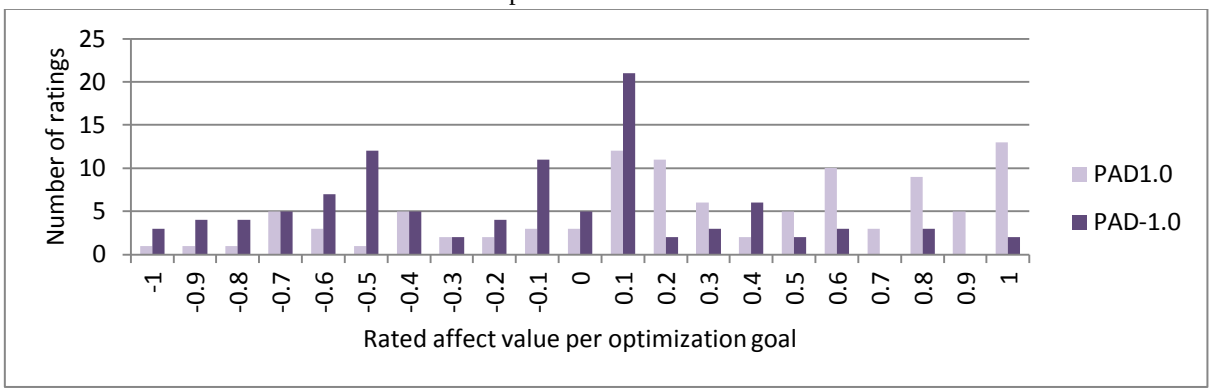

Figure 24. Histograms of affect ratings on the P dimension for 10 artworks optimized for PAD1.0 and 10 artworks optimized for PAD-1.0.

5. Discussion and future work

The paper presented a full system to generate affective abstract art based on the visual properties of existing human made art. The results of the experiments and the overall development of the full system raised opportunities for improvements and further research, and there were several limitations that severely influenced the effectiveness of the system to generate art with an intended affect. Both topics will be discussed in this section.

Measuring affect. The image features used to determine the affect of artworks were the most crucial limitation in this study. These features were not sufficient for capturing the relationship between the visual properties of an artwork and the rated affect. Possible reason for this is the fact that the used image features were bad representations of the way artworks are perceived by humans. Previous

studies suggest that features based on art principles and theory are better representations for this [18, 37], using these might thus result in more accurate measures to use in evolving affective artwork. But another reason could be that there actually is a poor or very complex relation between the visual properties of an artwork and the perceived affect, which is related to the subjectivity of the affect of abstract art. The low consistency in the ratings of the affect of both the original artworks as well as the generated artworks indicates that the affect of abstract art is highly personal. This might be a result of the use of the AffectButton as the method for affective self-report, since the AffectButton was not validated to be a reliable method to rate abstract art. But the low consistency in ratings also raises the question whether it is feasible to generate art that is generally perceived as having a certain (intended) affect. For example, the generated artworks that were at least a little effective in reaching their intended affect (the ones optimized for PAD 1.0 or PAD -1.0) showed great consistency between their visual properties. While this consistency is a consequence of the restricted style of the system and the poor affect measures, it could also be a result of the subjectivity of the affect of abstract art. Because these artworks were optimized towards extremities, this resulted in extremities in visual properties (very dark and empty versus very bright and busy). That these visually extreme artworks were rated closer to their intended affect is cause to believe that these extremities are actually necessary to reach consensus in the affect rating. For example, extreme visual properties leave less room for more subtle ones and the artworks are then perceived more similarly by all participants, lowering the subjectivity, increasing the consensus on affect. Since this research focussed purely on directly using a corpus of art as a method to retrieve affect measures, theory on art and theory on the affect of imagery was not included. Above observations show that including knowledge from these fields, like de Rooij's work [27], might aid in developing a more effective system for generating affective abstract art.
A point related to this is the relation between the perceived affect and the aesthetic value of an artwork. Although aesthetic value was not taken into account in this study, simply looking at the artworks that received generally more extreme affect ratings indicate that such a relation is likely. It is not hard to imagine that artworks that are perceived as more "beautiful" invoke stronger affect, or vice versa. Exploring this relation, either through existing studies or through new surveys that also measure aesthetic value, could result in interesting, combined measures for generating art.

Evolutionary setup. In the experiments on evolving art the choice was made to evolve for different affect goals. This meant that the parameters used in the evolutionary setup were kept the same throughout all experiments, which caused limitations on the exploration and exploitation within the optimization process. For example, there were no survivors selected in the process, meaning that offspring could have a lower fitness than its parents. Also, for tournament selection a size of 5 was used, which causes high selection pressure in a population size of 100. Experimenting with these parameters could result in better optimization on A and D or a higher variety in artworks optimized on P. Another possible enhancement of the evolutionary setup would be using multi-objective optimization to evolve for multiple dimensions, instead of one weighted objective as used in experiment 5. Furthermore, the experiments in section 4.3 showed the possibilities of manually adjusting the measures for affect and tweaking the parameters of individuals. It would be interesting to see if the affect measures and the generative system could be pushed further, possibly to reach the traditional goal of producing novel and aesthetically pleasing art.
Overall the development of the full end-to-end system yielded interesting results on the affect of abstract art, but it also gave insight into the process of using three continuous values as a basis to learn fitness functions. Where usually measures are created to evolve for one specific goal, like high aesthetic value, measures for continuous values provide the option to evolve for a very specific affect. Although the learned measures in this study were not accurate enough to be used to effectively evolve art for specific affect values, the ability to optimize for a set value could provide great flexibility in

generating art. It would be interesting to see how using accurate measures for continuous values as a fitness function in an evolutionary art setup would contribute to the ability to control the generated output of the system.

6. Conclusions

The primary goal of this study was to generate affective abstract art based on affect measures learned from an existing corpus of abstract art. This goal was partially reached. Affect measures were learned for the three affect dimensions pleasure, arousal and dominance. But these measures were determined to be poor, especially for measuring affect on the arousal and dominance dimensions. Due to these low accuracies, the study focussed on the affect dimension pleasure, including the other dimensions as secondary measures. Art was then generated in a restricted style through an evolutionary system. For this system a specific affect goal could be set on one or more dimensions, and the learned measures were then applied to determine the fitness of generated images based on the set affect goal. Using these calculated fitness values, the system was able to optimize artworks towards measured pleasure values of -1.0, 1.0 and as high as possible, resulting in sets of unique, but visually similar, artworks for each goal. Additional experiments determined that the measures could also be used to optimize for an affect goal on multiple dimensions. Results of a second web survey showed that the system was capable of generating artwork on a broad range of affect, but that artworks optimized on only the pleasure dimension were not rated as having their intended affect. Artworks optimized to the extremities on all three dimensions showed more promising results, being rated more closely to the affect value they were optimized for. So the developed system succeeded in learning measures for affect, but due to their low accuracy they were not effective fitness functions for generating affective art, resulting in a limited set of artworks that was actually perceived as having their intended affect.
The secondary goal of the study was the exploration of the process of using example-based learning to build prediction models to be directly applied as measures for evolving art. Despite the limitations of this study, the setup seems to be a viable method to determine fitness functions. Especially the ability to optimize for specific set values, by learning measures for continuous values, is an interesting opportunity provided by this specific setup.

References

1. Baluja, S., Pomerleau, D., & Jochem, T. (1994). Towards automated artificial evolution for computer-generated images. *Connection Science*, *6*(2-3), 325-354.
2. Birchfield, D. (2003). Generative model for the creation of musical emotion, meaning, and form. In *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence* (pp. 99-104). ACM.
3. Broekens, J., & Brinkman, W. P. (2009). Affectbutton: Towards a standard for dynamic affective user feedback. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (pp. 1-8). IEEE.
4. Correia, J., Machado, P., Romero, J., & Carballal, A. (2013). Feature selection and novelty in computational aesthetics. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design* (pp. 133-144). Springer Berlin Heidelberg.
5. Draves, S. (2005). The electric sheep screen-saver: A case study in aesthetic evolution. In *Applications of Evolutionary Computing* (pp. 458-467). Springer Berlin Heidelberg.
6. Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis* (Vol. 3). New York: Wiley.
7. Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion* 6(3-4), 169-200.

8. de Freitas, A. R., & Guimarães, F. G. (2011). Originality and diversity in the artificial evolution of melodies. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation* (pp. 419-426). ACM.
9. Galanter, P. (2003). What is Generative Art? Complexity theory as a context for art theory. In *GA2003–6th Generative Art Conference.*
10. Galanter, P. (2010). The problem with evolutionary art is… In *Applications of Evolutionary Computation* (pp. 321-330). Springer Berlin Heidelberg.
11. Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.
13. Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, (6), 610-621.
14. den Heijer, E., & Eiben, A. E. (2010). Using aesthetic measures to evolve art. In *IEEE Congress on Evolutionary Computation (CEC), 2010* (pp. 1-8). IEEE.
15. den Heijer, E., & Eiben, A. E. (2011). Evolving art using multiple aesthetic measures. In *Applications of Evolutionary Computation* (pp. 234-243). Springer Berlin Heidelberg.
16. Johnson, C. G. (2012). Fitness in evolutionary art and music: what has been used and what could be used?. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design* (pp. 129-140). Springer Berlin Heidelberg.
17. Lee, P., Teng, Y., & Hsiao, T. C. (2012). XCSF for prediction on emotion induced by image based on dimensional theory of emotion. In *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation* (pp. 375-382). ACM.
18. Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia* (pp. 83-92). ACM.
19. McCormack, J. (2005). Open problems in evolutionary music and art. In *Applications of Evolutionary Computing* (pp. 428-436). Springer Berlin Heidelberg.
20. Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14*(4), 261-292.
21. Meijering, E. (2014). FeatureJ: A Java Package for Image Feature Extraction [Computer software]. Retrieved from http://www.imagescience.org/meijering/software/featurej/
22. Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., & Reuter-Lorenz, P. A. (2005). Emotional category data on images from the International Affective Picture System. *Behavior research methods*, *37*(4), 626-630.
23. Phon-Amnuaisuk, S., Law, E., & Kuan, H. (2007). Evolving music generation with SOM-fitness genetic programming. In *Applications of Evolutionary Computation* (pp. 557-566). Springer Berlin Heidelberg.
24. Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion, 1*.
25. Quinlan, J. R. (1992, November). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343-348).
26. Ridler, T. W., & Calvard, S. (1978). Picture thresholding using an iterative selection method. *Systems, Man and Cybernetics, IEEE transactions on, 8*(8), 630-632.
27. de Rooij, A., Broekens, J., & Lamers, M. H. (2013). Abstract expressions of affect. *International Journal of Synthetic Emotions (IJSE), 4*(1), 1-31.
28. Russ, J. C., & Woods, R. P. (1995). The image processing handbook. *Journal of Computer Assisted Tomography, 19*(6), 979-981.

29. Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology, 39*(6), 1161.
30. Scirea, M., Nelson, M. J., & Togelius, J. (2015). Moody Music Generator: Characterising Control Parameters Using Crowdsourcing. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design* (pp. 200-211). Springer International Publishing.
31. Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. R. K. (2000). Improvements to the SMO algorithm for SVM regression. *Neural Networks, IEEE Transactions on, 11*(5), 1188-1193.
32. Takagi, H. (2001). Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE, 89*(9), 1275-1296.
33. Walker, R. F., Jackway, P., & Longstaff, I. D. (1995). Improving co-occurrence matrix feature discrimination. In *DICTA'95, 3rd Conference on Digital Image Computing: Techniques and Application* (pp. 643-648).
34. Wang, Y., & Witten, I. H. (1997). Induction of model trees for predicting continuous classes. *Poster papers of the 9th European Conference on Machine Learning, 1997*.
35. Yanulevskaya, V., Uijlings, J., Bruni, E., Sartori, A., Zamboni, E., Bacci, F., ... & Sebe, N. (2012). In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 349-358). ACM.
36. Zhang, H., Augilius, E., Honkela, T., Laaksonen, J., Gamper, H., & Alene, H. (2011). Analyzing emotional semantics of abstract art using low-level image features. In *Advances in Intelligent Data Analysis X* (pp. 413-423). Springer Berlin Heidelberg.
37. Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T. S., & Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *Proceedings of the ACM International Conference on Multimedia* (pp. 47-56). ACM.